УДК 81’33, 81.512.1

Р.А. Гыйльмуллин,
Б.Э. Хәкимов,
В.Р. Гафарова

ТАТАР ТЕЛЕНДӘГЕ ТЕКСТЛАРДА ИСЕМЛӘНГӘН ОБЪЕКТЛАРНЫ ТАМГАЛАУ МӘСЬӘЛӘСЕНӘ КАРАТА

Статья посвящена проблеме разметки именованных сущностей в текстах на татарском языке. Для татарского языка задача распознавания таких объектов в рамках автоматической обработки естественного языка находится на начальном этапе, по данному вопросу опубликовано несколько статей и экспериментальных моделей. Авторами создан размеченный корпус именованных сущностей на татарском языке, состоящий из предложений, отобранных случайным образом из различных текстов. Для корпуса была принята классификация, состоящая из 11 типов сущностей, разметка осуществлена без вложенных и пересекающихся объектов. Размеченный корпус может быть использован для машинного обучения и тестирования моделей распознавания именованных сущностей в текстах на татарском языке.

Ключевые слова: татарский язык, автоматическая обработка естественного языка, именованная сущность, ономастика, разметка.

This paper discusses the problem of named entity annotation in Tatar texts. For the Tatar language, the named entity recognition task is at an early stage, with a few articles and experimental models published. A tagged corpus of named entities in the Tatar language was developed. It contains sentences randomly selected from various sources. A classification of 11 named entity types was adopted for the corpus, while nested and intersecting entities have not been annotated. The Tatar named entity corpus can be used for machine learning and testing models for named entity recognition in Tatar texts.

Keywords: tatar language, natural language processing, named entity, onomastics, linguistic annotation.

Исемләнгән объектларны тану (рус. «распознавание именованных сущностей», инг. «named entity recognition») – табигый телләрне автомат рәвештә эшкәртү (инг. «natural language processing») өлкәсендәге стандарт биремнәрнең берсе. Хәзерге вакытта ул гадәттә нейрон челтәрләренә нигезләнгән махсус модельләр ярдәмендә башкарыла. Татар теле өчен бу мәсьәлә әлегә башлангыч дәрәҗәдә генә хәл ителгән. Кайбер алымнар һәм эксперименталь модельләр аз санлы фәнни мәкаләләрдә тасвирлана [Nevzorova et al, 2018] яки махсус репозиторийларда тәкъдим ителә [https://github.com/ksemiya/NER_in_Tatar].

Инглиз телендәге *«named entity»* оригиналь термины рус теленә *«именованная сущность»,* төрек теленә «adlandırılmış varlık» яки *«varlık ismi»* [Özger, Diri, 2016], казакъ теленә *«аталған нысан»*

(«нысан» – «объект») дип тәрҗемә ителә. Шуларны исәпкә алып, әлеге мәкаләдә *«исемләнгән объект»* термины тәкъдим ителә.

Исемләнгән объект – гамәли мәгълүмат эшкәртү белән бәйле төшенчә. Тел белемендә ул күбесенчә ялгызлык исемнәргә туры килә, әмма реаль чынбарлыктагы конкрет бер объектка (зат, шәхес, географик урын, оешма һ.б.) нисбәт ителергә тиеш була [Nouvel et al, 2016]. Шул күзлектән караганда, *Рәис – гарәп теленнән кергән исем* һәм *Биектау исемле авыллар республикабызда берничә* кебек җөмләләрдә ялгызлык исемнәр булса да, исемләнгән объектлар юк булып чыга, чөнки конкрет референт аталмый.

Текстларны автоматик эшкәртү өлкәсендә гомум кабул ителгәнчә, текстларда фактларны автомат рәвештә ачыклау нәтиҗәлерәк булсын өчен, ялгызлык исемнәр белән белдерелгәннәрдән тыш, дата, вакыт, акча һәм башка үлчәү берәмлекләрен эченә алган микъдари гыйбарәләр дә исемләнгән объектлар буларак карала [Nadeau, 2007]. Шулай итеп, андый объектларны белдерә торган тел чараларына антропонимнар, топонимнар, эргонимнар һәм башка ялгызлык исемнәр, шулай ук даталарны, вакытны, төрле валюталарда акча күләмен, һәртөрле физик зурлыкларны һ.б. аңлата торган санлы сүзтезмәләр керә.

Исемләнгән объектларны тану модельләрен булдыру һәм өйрәтү өчен тиешле тел берәмлекләре тамгаланган мәгълүмат тупланмалары (корпуслар) кулланыла. Шул максаттан, ТР Фәннәр академиясенең Гамәли семиотика институтында татар телендәге текстларда исемләнгән объектларны тану биремнәрен башкару өчен, 5800 җөмләдән торган махсус тамгаланган корпус төзелде. Аерым җөмләләр төрле чыганаклардан алынды, автоматик ысул белән баш хәрефтән язылган сүзләр очраган мисаллар сайланды. Корпусны тамгалау өчен Label инструментарие базасында А.Ф. Хөсәенов тарафыннан эшләгән махсус веб-интерфейс файдаланылды [https://labelstud.io/].

Куелган максатларга бәйле рәвештә исемләнгән объектларның төрле классификацияләре кулланылырга мөмкин, аерып чыгарыла торган категорияләр саны исә өч-дүрттән алып берничә дистәгә җитә ала [https://grobid-ner.readthedocs.io/en/latest/class-and-senses/]. Аерым объектларның бер җөмләдә бер-берсенә мөнәсәбәтенә карап, төрле тамгалау ысуллары гамәлгә ашырыла. Мәсәлән, Dialogue Evaluation бәйгесе өчен OpenCorpora платформасында тәкъдим ителгәнчә, берничә дәрәҗәле тамгалау ысулы белән эшләп була [Starostin, A. et al, 2016]. Моннан тыш, бер-берсе эченә кергән һәм чикләре кисешкән объектларны исәпкә алмыйча, һәр берәмлеккә бер генә категорияне билгеләүдән гыйбарәт булган ысул да бар [Можарова, Лукашевич, 2016]. Шулай ук, исемләнгән объектларны тамгалаганда телләрнең спецификасын һәм милли-мәдәни үзенчәлекләрне дә игътибарга алу мөһим.

Әзерләнгән корпуста исемләнгән объектлар 11 төргә бүлеп, кисешүләрсез һәм бер-берсе эченә кертелмичә тамгаланды. Корпуста түбәндәге шартлы билгеләр кулланыла: PERSON – кешеләр белән бәйле ялгызлык исемнәре; TITLE – вазифалар, мактаулы исемнәр,

титуллар һ.б.; ORGANIZATION – оешмалар, учреждениеләр һ.б.; LOCATION – географик урыннар; DATE – даталар; TIME – вакыт аралыклары; MONEY – акча берәмлекләре; QUANTITY – теләсә нинди үлчәнешләр (үлчәү берәмлекләре белән бирелгән саннар); CARDINAL – башка төр саннар; FACILITY – аерым корылмалар һәм төзелеш объектлары; OTHER – башка төрдәге объектлар.

1. **PERSON** категориясенә реаль яки уйдырма кеше исемнәре, фамилияләр, ата исемнәре (шул исәптән кызы, улы / углы компонентлары белән ясалган очраклар), кушаматлар, әдәби әсәрләрдәге персонаж исемнәре кертелде.

Татар теленең үзенчәлеге буларак, бу категориядә кешене атый торган ялгызлык исемнәре белән берлектә туганлык атамалары һәм традицион мөрәҗәгать итү сүзләре дә тамгаланды: *апа, абый, түти, дәдәй, ага, бабай, әби, карчык, карт, әфәнде, ханым, туташ, хәзрәт.* Мәсәлән: ***Нурихан ага Фәттахның (PERSON)*** *хатыны* ***Руфина апа (PERSON)*** *да чакырып шалтыраткач, кызыксынуым тагын да көчәйде.*

Хәзерге вакытта продуктив булмаган, әмма тарихи, дини һәм башка текстларда очрый торган, гарәп үрнәге буенча *бинт, бинте*, *ибн, ибне, бин* компонентлары белән ясалган модельләр дә әлеге категориядә билгеләп үтелде: *Шуһабетдин бине Баһаветдин, Мөхлисә бинт Габделгалләм.* Мәсәлән: ***Галимҗан Әл-Баруди (PERSON) (Галимҗан бине Мөхәммәдҗан бине Бинеямин бине Гали бине Колмөхәммәд) (PERSON)*** *1857 елның 2 февралендә (DATE) Казан губернасы (LOCATION) Казан өязе (LOCATION) Чуашиле авылында (LOCATION) туган.*

Текстта бер төркем кешеләр уртак исем белән бәйләнгән очракларда бөтен фрагмент бер элемент буларак билгеләнде: ***Дилия белән Зилә Нурмөхәммәтовалар (PERSON)****.*

Оешма, учреждение һәм корылмаларның атамалары составында кеше исемнәре һәм фамилияләре булган очраклар, исемләнгән объектның гомуми мәгънәсеннән чыгып, бердәм элемент буларак тасвирланды. Мәсәлән, *Г. Тукай исемендәге аэропорт, В.И. Ленин музей-йорты, В.И. Ленин исемендәге Казан дәүләт университеты* кебек мисаллар FACILITY (корылмалар һәм төзелеш объектлары) яки ORGANIZATION (оешмалар) категориясенә, кертелде: *Әйтик, 2003 елда (DATE)* ***Габдулла Кариев исемендәге Яшь тамашачы театры (ORGANIZATION)*** *Рабит Батулла (PERSON) пьесасы буенча «Сак-сок» драма-бәетен (OTHER) чыгарган иде.*

Куштырнаклар эчендә бирелгән ялгызлык исемнәр (топонимнар һәм антропонимнар), оешма атамалары һ.б. исемләнгән объектның төре нигезендә тамгаланды, мәсәлән, *«Мәскәү» кунакханәсе – FACILITY.* Шулай ук: *Ә без* ***«Зөләйха күзләрен ача»ны (OTHER)*** *тәрҗемә иткәне өчен, күзен дә ачырмыйча тиргәдек кенә аны.*

2. Үз чиратында, **TITLE** категориясенә вазифа атамалары, мактаулы исемнәр, шәхескә йөкләнә торган функцияләрнең исемнәре

керде. Уникаль вазифа атамалары үзләре бәйләнгән локация яки оешма исеме белән бергә алынды. Мәсәлән: *Соңрак концерт мәйданында республика сайлаучыларын* ***Татарстан Президенты (TITLE)*** *Рөстәм Миңнеханов (PERSON) сәламләячәк. Оешма каршында ачылган һәйкәлдә –* ***Советлар Союзы Герое (TITLE), полковник-элемтәче (TITLE)*** *Борис Кузнецов (PERSON) образы гәүдәләндерелгән.*

3. **ORGANIZATION** категориясендә кешеләрнең эш урыннары яки башка коллектив эшчәнлеге белән бәйле булган исемләнгән объектлар урын алды. Бу категориягә оешма, компания, фирма, партия, ансамбль, команда, клуб атамалары; кыскартылган атамалар *(ЭЭМ, КПРФ)*; юридик зат төрләре *(ААҖ, ЯАҖ, РайПО)*; ялгызлык исемнәр белән белдерелгән оешма атамалары; билгеле бер территориядә урнашкан хөкүмәт, дәүләт, дәүләт хакимияте органнары атамалары *(Татарстанның данын яклау)*; хәзерге вакытта таркалган, яшәешен туктаткан яки башка исем астында йөргән илләр һәм дәүләт атамалары кертелде *(Идел буе Болгар дәүләте, Алтын Урда, Казан ханлыгы, Советлар Союзы)*. Мәсәлән: *Ни өчен* ***Мәскәү (ORGANIZATION)*** *белән* ***Грозный (ORGANIZATION)*** *арасындагы килешү барып чыкмаган? Закон нигезендә, бу өстәмә түләүләр* ***Россия Пенсия фондына (ORGANIZATION)*** *күчкән иминият кертемнәреннән чыгып исәпләнелә. 2005 елда (DATE)* ***«КАМАЗ» ААҖнең (ORGANIZATION)*** *фаразланган керем күләме 52 миллиард сум (MONEY) тәшкил итәчәк. Без ары таба да биографиябезне, мирасыбызны,* ***Советлар Союзы (ORGANIZATION)*** *балалары икәнлегебезне горурлык белән хәтердә тотачакбыз.*

4. Локация буларак (**LOCATION**) чынбарлыктагы урынны һәм хәрәкәт юнәлешен күрсәтә торган исемләнгән объектлар тамгаланды: географик объект атамалары; шәһәр, төбәк, урам исемнәре, йорт һәм / яки фатир номерлары, таулар, күлләр, диңгезләр, океаннар атамалары һ.б. Топонимның ыру төшенчәсен белдергән сүзләр дә исемләнгән объект составында билгеләнде (*өлкә, шәһәр, бистә, урам, йорт, елга* һ.б.). Мәсәлән: *Моннан берничә атна элек ремонтка ябылган күпер аркасында,* ***Яшел Үзән шәһәре (LOCATION)*** *белән* ***Мирный бистәсе (LOCATION)*** *арасын үтү, транспорт бөкеләре аркасында шактый проблемалыга әверелгән иде.*

Бүтән объектларның өлеше булган географик объектлар аерым тамгаланды: *Фатих Кәрим (PERSON) 1909 елның 9 гыйнварында (DATE)* ***Башкортстанның (LOCATION) Бишбүләк районында (LOCATION)*** *туа.*

Берничә элементтан торган төгәл адреслар бер исемләнгән объект буларак билгеләнде: ***Декабристлар ур., 83 (LOCATION)*** *адресындагы йортка киоск ялганган.*

5. **DATE** категориясенә төрле форматта бирелгән даталар һәм тәүлектән зуррак вакыт аралыклары кертелде: *20.01.2001, 2009 елның сентябре, 1934 елның 1 мае, 1990 еллар, 2000 еллар, 8–14 октябрь 2020, 1895–1919, 965 ел (б.э.к.) һ.б.* Мәсәлән: *Искәртеп узабыз,*

2016 ел (DATE) *өчен милек салымын түләү* ***1 декабрьгә (DATE)*** *кадәр башкарылырга тиеш.*

Корпуста һижри ай исемнәре дә тамгаланды: *Изге* ***Рамазан (DATE)*** *аенда Аллаһы Тәгаләгә ышанучылар саваплы гамәлләрне күбрәк башкара.*

6. **TIME** категориясендә тәүлектән кимрәк вакыт аралыклары, булган очракта, *сәгать, минут, секунд* сүзләре белән бергә тамгаланды. Мәсәлән: *Без барып җиткәндә* ***10 (TIME)*** *була, эшли башларга да өлгермибез, төшке аш җитә. Корабның радардан юкка чыгуы турында Мисыр хакимияте (ORGANIZATION) җомга көнне* ***13.00 сәгатьтә (TIME)*** *хәбәр иткән.*

7. **MONEY** категориясендә берәмлек исемнәрен дә кертеп акча күләмнәре тамгаланды: *2 миллиард доллар, 1 млн 400 мең сум.* Мәсәлән: *Җыелма бюджетта белем бирү өлкәсенә –* ***59,9 млрд (MONEY).***

8. **QUANTITY** категориясенә төрле үлчәү берәмлекләре (авырлык, аралык, тизлек, озынлык, мәйдан һ.б.) белән бирелгән саннар кертелде. Саннар үлчәү берәмлекләренең (метр, литр, тонна) атамалары белән, аларның кыскартылган тамгалары (м, кг, ц) һәм махсус тамгалары (мәсәлән, «°») белән бергә тамгаланды. Мәсәлән: *Менә шушы* ***114 гектарлы (QUANTITY)*** *басудан көзге арыш гектарыннан уртача* ***28,2 центнер (QUANTITY)*** *төшем белән куандырган. Туфракка* ***5 мм (QUANTITY)*** *күмдерергә кирәк. Аның параметрлары –* ***90-62-89 (QUANTITY)****, сылулыкка ул көндәлек хезмәт аша ирешә.*

Катлаулы исемләнгән объектлар берничә үлчәү берәмлеге белән белдерелгән очракларда, алар бер элемент (бер бөтен) буларак билгеләнде, мәсәлән: *3 м 50 см.*

9. **CARDINAL** категориясенә акча берәмлекләренә, даталарга, вакыт һәм үлчәү берәмлекләренә туры килмәгән башка саннар кертелде. Бу саннар цифрлар яки сүз белән бирелеп, саналмыш сүзләрдән башка тамгаланды: *Бүген шәһәрдә* ***1224 (CARDINAL)*** *урынга исәпләнгән* ***ике (CARDINAL)*** *яңа мәктәп төзелә.*

10. FACILITY категориясенә аерым корылмалар, төзелеш объектлары, парк, күпер һәм магистраль атамалары һ.б. кертелде: *Казан Кремле, Казан Богородица чиркәве, Г.Тукай исемендәге аэропорт, «Имәнлек» станциясе, Муса Җәлил исемендәге Татар опера һәм балет театры, «Яшьлек» паркы*. Мәсәлән: ***В.И. Ленин музей-йортын (FACILITY)*** *караргa килүче халык саны да 3 тапкырга артыр дип уйланыла.*

11. OTHER категориясенә башка төр объектлар: вакыйга атамалары (*Бөек Ватан сугышы*), күк җисемнәре, табигать күренешләре атамалары; китаплар, җырлар, конкурслар, фестивальләр, премияләр, бәйрәмнәр һәм истәлекле көннәрнең атамалары, сәнгать әсәрләре исемнәре; закон актлары атамалары, процент тамгалары белән бирелгән саннар, телефон номерлары һ.б. кертелде. Мәсәлән: *Наил*

Нәбиуллин (PERSON) узган елда ***«Азатлык» дигән гәзит (OTHER)*** *чыга башлауны хәбәр итеп, аның гомуммилли бәйсез басма булачагын әйтте. Аларда укучыларның* ***85 проценты (OTHER)*** *– инглизләр, чит ил балалары санын* ***15 проценттан (OTHER)*** *да арттырмыйлар. Ул Чабаксарда (LOCATION) узды, анда* ***«Россия – спорт державасы» халыкара форумы (OTHER)*** *бара. Бүген шәһәр-районнарда* ***Сабан туйлары (OTHER)*** *узды.*

Йомгак ясап әйткәндә, тупланган корпусны татар телендәге текстларда исемләнгән объектларны тану модельләрен булдыру һәм сынау өчен кулланырга мөмкин. Аерым очракларны тамгалауга карата тәкъдим ителгән алымнар бәхәсле була алса да, тикшеренү нәтиҗәләре мондый берәмлекләрнең категорияләрен билгеләү өлкәсендә бердәм карашлар булдыруга ярдәм итәр дип уйларга нигез бар.

Әдәбият

Инструкция по разметке сущностей для Dialogue Evaluation 2016. URL: http://opencorpora.org/wiki/Nermanual/2/model.

Можарова В.А., Лукашевич Н.В. Двухэтапный подход к извлечению именованных сущностей // Труды конференции по искусственному интеллекту КИИ-2016. Т. 2., 2016. С. 81–88.

GROBID NER Named entity classes, URL: https://grobid-ner.readthedocs.io/en/latest/class-and-senses/

Nadeau D., Satoshi S. A survey of named entity recognition and classification // Lingvisticae Investigationes. 2007. № 30. P. 3–26.

NER_in_Tatar, URL: https://github.com/ksemiya/NER_in_Tatar.

Nevzorova O., Mukhamedshin D., Galieva A. Named Entity Recognition in Tatar: Corpus Based Algorithm // Proceedings of Computational Models in Language and Speech Workshop (CMLS 2018). CEUR. Vol. 2303. P. 58–68.

Nouvel, D., Ehrmann, M., & Rosset, S. Named Entities for Computational Linguistics. New York: Wiley, 2016. 192 p.

Open Source Data Labeling, URL: https://labelstud.io/.

Özger, Z.B., & Diri, B. Türkçe Dokümanlar İçin Kural Tabanlı Varlık İsmi Tanıma // Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi. Vol. 5, No. 2. Jun. 2016.

Starostin, A. et al. FactRuEval 2016: Evaluation of Named Entity Recognition and Fact Extraction Systems for Russian // Компьютерная лингвистика и интеллектуальные технологии: Вып. 15 (22). М.: Изд-во РГГУ, 2016. С. 688–705.

Гыйльмуллин Ринат Абрек улы,
физика-математика фәннәре кандидаты,
ТР ФА Гамәли семиотика институты директоры

Хәкимов Булат Эрнст улы,
филология фәннәре кандидаты, Казан федераль университетының билингваль һәм цифрлы белем бирү кафедрасы доценты,
ТР ФА Гамәли семиотика институтының әйдәп баручы фәнни хезмәткәре

Гафарова Вилүзә Роберт кызы,
филология фәннәре кандидаты, ТР ФА Гамәли семиотика институтының әйдәп баручы фәнни хезмәткәре