Eestikeelse küsimus-vastus süsteemi arendamine

Magistritöö laiendatud kokkuvõte

Anu Käver (182912IAPM), juhendaja Tanel Alumäe

1.Teema kirjeldus, valiku põhjendus

Oluline osa tehisintelligentsi arendamisest on loomuliku keele töötlemine. Selle eesmärgiks on töötada välja arvutisüsteemid, mis suudavad käsitleda inimkeeli samaväärselt või paremini, võrreldes inimesega – teksti ja kõnet genereerida, tõlkida, mõista jne (1). Keele mõistmise kategoorias on üheks oluliseks eesmärgiks sellised süsteemid, mis suudavad etteantud tekstidest leida informatsiooni vastusena inimese poolt loomulikus keelepruugis esitatud küsimustele. Selliste küsimus-vastus süsteemide kasutusvaldkond hõlmab otsingumootoreid, juturoboteid, dialoogisüsteeme jne.

Tänapäeval on küsimus-vastus süsteemide nagu ka enamiku teiste loomuliku keele valdkonna ülesannete puhul saavutatud suurimat edu närvivõrkude abil (2). Seejuures on paremad tulemused saavutatud ülisuurte andmemahtude peal treenitud kontekstipõhiste[1] keelemudelitega (3) (4), mida saab täpsemini edasi treenida vastavalt püstitatud eesmärgile (nt kõne genereerimine, tõlkimine, lauseosade tuvastamine, küsimustele vastamine). Kontekstipõhine keelemudel hõlmab endas keele sõnavara vektorite kujul kõrge dimensionaalsusega vektorruumis (sadu dimensioone) (5). Sellisel viisil on sõnade omavahelised seosed kogu keeles esitatud matemaatiliselt ja need seosed on aluseks mudeli edasisel treenimisel täpsema ülesande kontekstis.

Eesti keeles on olemas üks suur kontekstipõhine keelemudel – EstBERT, mis on avaldatud novembris 2020 (6). Selle treenimisel võeti aluseks ingliskeelse BERTi (5) (7) koostamise metoodika. Samuti on eesti keel olnud kasutusel paari suure mitmekeelse mudeli koostamisel (mBERT (8) ja XLM-RoBERTa (9) (10)), mille puhul treeniti sadakonda keelt paralleelselt, otsekui üht, kõigi nende keelte sõnavara sisaldavat keelt (11).

Olemasolevate eesti- ja mitmekeelsete mudelite baasil on proovitud eesti keeles lahendada mitmeid loomuliku keele ülesandeid, näiteks pärisnimede äratundmist, lauseosade määramist, morfoloogilisi ülesandeid, tekstide teema ja nende meelestatuse klassifitseerimist (*sentiment analysis*) (12). Teadaolevalt ei ole proovitud sedasama veel küsimus-vastus süsteemiga.

Küsimus-vastus süsteemi puhul on oluline ka andmestik, millest lähtuvalt aluseks olevat universaalset keelemudelit täpsemini treenida ja testida – ehk suures koguses küsimusi ja vastuseid. Eesti keeles puudub teadaolevalt ka selline andmestik.

Inglise keeles on väga tuntud küsimus-vastus andmestik SQUAD (13), millest on võetud eeskuju ka teiste keelte puhul (14). Tegu on Wikipedial baseeruva andmestikuga, mis koosneb tekstilõikudest ja selle kohta esitatud küsimustest, kus vastuseks on lühim küsimusele vastav osa tekstilõigust. Andmestik hõlmab esimeses versioonis umbes 100 000

[1] Mudeli treenimisel on pööratud tähelepanu iga sõna kontekstile treeningtekstides

küsimust, teises versioonis on lisatud umbes 50 000 küsimust, millele vastamine pole konteksti põhjal võimalik (15). Wikipedial põhineb teisigi suuri andmestikke (16), nt WikiQA (17), mis kasutab otsimootorite päringuid. SQUADi puhul on aga küsimused-vastused spetsiaalselt selle ülesande jaoks reaalsete inimeste poolt koostatud.

Viimastel aastatel on eri riikides tehtud mitmeid eksperimente, kuidas arendada küsimus-vastus süsteeme olukorras, kus vaatluse all olevas keeles napib ressursse – puudub kas suur keelemudel, või piisavalt suur treening- ja testandmestik küsimuste-vastuste kujul (18) (19) (20) (21) (11). On vastavaid teistes keeltes olemasolevaid ressursse sihtkeele hüvanguks ära kasutatud (*transfer learning*). Nt on töö eri etappides masintõlget rakendatud, aga ka mitmekeelset mudelit ilma tõlkimata kasutatud (nn *zero-shot* metoodika, kus mudelit treenitakse keeles, kus on olemas suur andmestik, ja treenitud mudel omandab võimekuse vastata küsimustele ka väikeste ressurssidega sihtkeeles). Selliseid eksperimente pole teadaolevalt eesti keeles tehtud.

2. Kirjanduse ülevaade

Heaks näidiseks erinevatest meetoditest, mille abil on uurijad viimastel aastatel küsimus-vastus süsteemide puhul keeltevahelist masinõpet rakendanud, on Jiahua Liu ja teiste Pekingi ülikooli egiidi all avaldatud **„XQA: A Cross-lingual Open-domain Question Answering Dataset“** (18).

Autorid realiseerivad kolm eri lähenemist küsimus-vastus süsteemidele ja rakendavad neid kaheksa suhteliselt väheste ressurssidega keele puhul.

Esimese kahe lähenemise puhul on aluseks ühekeelsed kontekstuaalsed mudelid. Rakendatakse masintõlget ja seda kahel viisil – esiteks, tõlgitakse ingliskeelsed treeningandmed sihtkeelde, ja treenitakse sihtkeelne mudel. Teiseks, treenitakse ingliskeelne mudel, ning seejärel tõlgitakse sihtkeele testandmed testimiseks inglise keelde. Autorid nendivad, et selline lähenemine sõltub tugevalt masintõlke kvaliteedist, eeskätt kannatab pärisnimede tõlkimine. Näiteks tõid autorid ühe Hiina lossi nime, mis tõlkes oli arusaamatul põhjusel asendatud teise hiinakeelse nimega. Samas on pärisnimed sageli ülioluline osa küsimustele vastuste leidmisel.

Kolmanda lähenemise puhul võeti aluseks mitmekeelne BERT, treeniti seda ingliskeelsete andmetega ja testiti tulemust sihtkeeles (*zero-shot* metoodika).

Selgus, et kõigis keeltes (välja arvatud võrdluseks sisse võetud inglise keel) saadi parimad tulemused mitmekeelse BERTi abil. Tulemuste F1-skoor[2] varieerus 13-39% vahel. Autorid põhjendavad suhteliselt viletsaid tulemusi sellega, et keeltevaheline küsimus-vastus süsteemide treenimine ongi keeruline ülesanne. Samas on näha teiste analoogsete uurimustööde puhul ka paremaid tulemusi, näiteks järgmises analüüsitavas artiklis.

Zero-shot lähenemist võrrelduna masintõlkega on kasutatud ka Facebook AI ja Londoni Ülikooli teadlaste, Patrick Lewise ja teiste poolt artiklis **„MLQA: Evaluating Cross-lingual Extractive Question Answering**“ (21).

[2] Laialtlevinud mõõdik küsimus-vastus süsteemide soorituse hindamiseks. Mõõdab sõnade kattuvust ennustatud vastuses ja igas tõesena märgistatud vastuses. Võtab iga küsimuse kohta aluseks kõige parema skooriga vastusevariandi. Lõpuks leitakse kogu testandmestiku kohta aritmeetiline keskmine. F1-skoori maksimaalne väärtus on 100%.

Nende lähenemine on eelmise artikli omast osaliselt erinev. Nad keskendusid testandmete koostamisel (treeningandmed olid vastavalt *zero-shot* metoodikale ingliskeelsed, pärit SQUAD-ist) sellele, et kõigis neid huvitanud kuues keeles ja ka inglise keeles oleks olemas paralleelsed andmed. See tähendab, et oleks olemas identse sisuga Wikipedia artiklid, mille puhul algselt koostatakse küsimused-vastused ingliskeelse sisu põhjal ja seejärel tõlgitakse teistesse keeltesse. Tulemusena peaks olema andmestikud ja neil saadud treeningtulemused hästi võrreldavad.

Lisaks *zero-shot* metoodikale kasutati ka siin mõlemat masintõlke lähenemist – enne treenimist SQUADi andmete tõlkimist sihtkeelde, ja testandmete tõlkimist inglise keelde.

Treenimisel võeti iga lähenemise puhul aluseks kolm mudelit – ingliskeelne BERT, mitmekeelne BERT ja mitmekeelne XLM mudel.

Tulemused olid mitmekesisemad – pooltes keeltes andis parimaid tulemusi *zero-shot* lähenemine XLM keelemudelil, aga teistes treeningandmete tõlkimine sihtkeelde ja selle andmestikuga mitmekeelse BERTi treenimine. Tulemused olid esimesena analüüsitud hiina autorite omast paremad, F1-skoor varieerus vahemikus 54-68%. Erinevuse üheks põhjuseks võib olla rangelt paralleelne andmestik, identsete andmete olemasolu suurte ja väikeste ressurssidega keeltes. Tekib küsimus, millise skoori saavutaks mudel, kui sellist kunstlikult paralleelset andmestikku ei oleks.

Minu töö kontekstis on väga olulised Tartu Ülikooli teadlaste Claudia Kittaski, Kairit Sirtsi ja teiste artiklid **„Evaluating multilingual BERT for Estonian**“ (12) ja **„EstBERT: A Pretrained Language-Specific BERT for Estonian“** (6). Esimene neist kasutab samuti mitmekeelseid mudeleid erinevates eesti keele loomuliku keele ülesannetes, teine tugineb võrdlusandmete hankimiseks esimesele, aga pakub sinna kõrvale suure eestikeelse BERTi eeskujul koostatud keelemudeli ja võrdleb selle abil saadud tulemusi.

EstBERT on koostatud u 1,1 miljardist sõnast koosneva treeningkorpuse abil, samas kui BERTi puhul oli mahuks u 3,3 miljardit sõna (5). Seega on andmemahud võrreldavad ja samas suurusjärgus.

Autorid lahendavad esimeses artiklis kuut ülesannet (kahel eri viisil lauseosade leidmist, morfoloogilist sildistamist, teksti klassifitseerimist, teksti meelsuse tuvastamist, pärisnimede äratundmist). Mitmekeelsed mudelid, mille alusel nad neid ülesandeid lahendasid, olid mBERT, DistilBERT (BERTi kontsentreeritud vorm), XLM-100 ja XLM-RoBERTa. Kõiki mudeleid treeniti erinevalt eelnevalt analüüsitud artiklitest sihtkeele ehk eesti keele korpustel. Kõige paremaid tulemusi andis XLM-RoBERTa mudel.

EstBERTi publitseerimisel võtsid autorid aluseks mitmekeelsete mudelite tulemused ja lahendasid samad kuus ülesannet ka EstBERTil. Autorite hüpotees, mida nad väljendasid ka juba esimeses artiklis, oli, et mitmete teiste keelte puhul on omakeelne suur keelemudel andnud paremaid tulemusi kui mitmekeelsed mudelid. Hüpotees leidis tõestust ning kuuest ülesandest viie puhul olid EstBERTil põhinevad tulemused parimad. Ainult teksti meelsuse tuvastamisel (*sentiment classification*) võimaldas XLM-RoBERTa paremaid tulemusi.

Autorite järeldus on, et järgmise sammuna võiks proovida eestikeelse RoBERTa arendamist. (RoBERTa mudel on BERTile teiste autorite poolt tehtud edasiarendus, kus mudelit on optimeeritud hüperparameetrite muutmise abil ja mudelit on kauem treenitud (22).)

Küsimus-vastus süsteemi Eesti autorid neis töödes realiseerinud ei ole.

3. Probleemi püstitus

Eelneva taustal on minu magistritöö eesmärgiks arendada välja eestikeelne küsimus-vastus süsteem. Selline süsteem hetkel eesti keele jaoks puudub ning see oleks aluseks järgmistele uurimistöödele. Ühtlasi koostan ma hetkel puuduva eestikeelse küsimus-vastus andmestiku, mida samuti on võimalik hiljem täiendada ja laiendada.

Võtan andmestiku koostamisel aluseks eestikeelse Wikipedia, nii et samast valdkonnast on võtta ka SQUAD ja vajadusel teisigi ingliskeelseid andmestikke, ja ühtsetel alustel mudeleid treenida ja testida.

Plaanis on proovida läbi mitu erinevat lähenemist, mida on kasutatud ka eelnevalt analüüsitud artiklites ja mida täpsemalt kirjeldab metoodika osa. Eesmärk on uurida, milline neist töötab eesti keele puhul kõige paremini. Hüpoteesiks on, et tulemusi mõjutavad nii treeningandmestiku suurus kui ka eesti keele osakaal aluseks olevas keelemudelis (mitmekeelsete mudelite puhul on see väike).

Arvuliseks eesmärgiks on koostada mudel, mille F1-skoor ulatub üle 60% ehk vahemikku, mille saavutasid minu poolt viidatud teise artikli autorid. Tegu pole väikese skooriga, kuna SQUADi esimese versiooni koostajad proovisid läbi ka juhuvalikuga küsimustele vastamise, mille F1-skooriks tuli 4,3% ja täpsete vastuste osakaaluks 1,3%. Tegu on selgelt keerulisema ülesandega kui näiteks lauseosade kaardistamine, kus võimalikke variante on iga sõna kohta kindlalt ette antud valikus. Küsimusele vastates võib teoreetiliselt valida vastuseks ükskõik millise teksti alamlõigu ehk üle 60% ulatuv täpsus oleks hea tulemus.

4. Metoodika valik ja põhjendus

4.1. Mudelid

Võttes eeskuju tutvustatud meetoditest, mida on kasutanud teised autorid, on plaanis koostada järgmised küsimus-vastus mudelid:

1. Aluseks mitmekeelne XLM-RoBERTa, treeninguks ingliskeelne SQUAD andmestik, testimiseks minu oma koostatud uus eestikeelne andmestik (*zero-shot*)
2. Aluseks mitmekeelne mBERT, treeninguks ingliskeelne SQUAD andmestik, testimiseks minu oma koostatud uus eestikeelne andmestik (*zero-shot*)

3. Aluseks mitmekeelne XLM-RoBERTa, treeninguks masintõlkega eesti keelde tõlgitud SQUAD andmestik, testimiseks minu oma koostatud uus eestikeelne andmestik
4. Aluseks mitmekeelne mBERT, treeninguks masintõlkega eesti keelde tõlgitud SQUAD andmestik, testimiseks minu oma koostatud uus eestikeelne andmestik

5. Aluseks eestikeelne EstBERT, treeninguks masintõlkega eesti keelde tõlgitud SQUAD andmestik, testimiseks minu oma koostatud uus eestikeelne andmestik

Kokkuvõtlikult kavatsen kasutada kolme erinevat alusmudelit ja läheneda selle treenimisele kahel erineval viisil – ingliskeelse andmestikuga, või tõlkides sama andmestiku eesti keelde.

Eelnevatest artiklitest tuli välja, et masintõlke kasutamine küsimus-vastus süsteemi puhul võib olla problemaatiline. Tekkida võivad tõlkevead, tõlkimissüsteem võib käsitleda valesti pärisnimesid, mis on küsimustele vastamisel väga olulised. Seda tuleb võtta tulemuste hindamisel arvesse.

Mõlema lähenemise puhul treeningandmetele otsin ma tegelikult (ja otsisid ka eeskujuks võetud autorid) lahendust olukorrale, kus omakeelseid treeningandmeid ei ole ja selle asemel proovitakse kasutada ingliskeelseid. Siit tekib kohe ka küsimus, kas on kuidagi võimalik lahendada juurpõhjus ja eestikeelsed treeningandmed siiski tekitada. Vajalikud andmekogused oleksid siiski väga suured (SQUADi esimene versioon sisaldab 100 000 küsimust). Seega ei suudaks üks inimene seda teha ja vaja oleks organiseerida vabatahtlik või tasustatud „kampaania“. Selline otsus on võimalik, aga hetkel oleks sellega alustada ennatlik. Minu poolt näidisandmestiku kogumine (vt 4.2) on näidanud, et selles protsessis on nii mõndagi subjektiivset, minu tehtud tööd tuleks esmalt hinnata ja analüüsida, enne kui kaasata teisi inimesi.

Küll on siin võimalik kompromiss ehk minu kogutud andmed võib siiski jagada kaheks, lisades osa neist treeningandmetele ja teise abil testida. Selle eesmärgi jaoks võib kogutud andmestikku suurendada piiratud mahus, nii et ei pea teisi inimesi (suurel arvul) kaasama.

4.2. Andmestik

Laiendatud kokkuvõtte valmimise hetkeks on esmane andmestik juba koostatud, võttes aluseks metoodika, mida kasutati SQUADi koostamisel ning ka SQUADi eeskujul prantsuskeelse PIAF-andmestiku koostamisel (14).

Kasutatud on Wikipedia poolt perioodiliselt iga keele kohta agregeeritavat artiklite baasi (23) ning rakendatud sellele Project Nayuki PageRank algoritmi (24), mida kasutasid nii SQUAD kui ka PIAF. See algoritm võimaldab leida kõige relevantsemaid dokumente, lähtudes nende omavahelisest viidete struktuurist. Valisin välja eestikeelse Wikipedia esimesed 10 000 artiklit.

Järgnevalt loobusin kõigist artiklitest, mis koosnesid ainult loeteludest. Järgmiseks jätsin alles ainult artiklid, milles sisaldus vähemalt viis vähemalt 500 tähemärgi pikkust lõiku. Nende artiklite arvuks kujunes 746.

Järgnevalt valisin juhuslikkuse alusel välja 20 artiklit (võrdluseks, SQUADi andmestik põhineb 536 artiklil ehk erinevate artiklite või teemade arv on ka seal suhteliselt väike), ja arvestasin neis ainult vähemalt 500-tähemärgiste lõikudega. Lõike tuli kokku 226.

Kasutasin küsimuste ja vastuste koostamiseks tööriista cdQA-annotator (25), mis on sobilik ühe autori poolt lokaalses arvutis töötamiseks ja salvestab tulemused SQUADi andmestiku formaadis. Esitasin iga lõigu kohta viis küsimust ja iga küsimuse kohta leidsin lühima sellele vastava tekstikatke (SQUADi ja PIAFi puhul tehti samamoodi). Täpsemalt proovisin leida lausa mitu varianti lühimast tekstikatkest, proovides imiteerida eri inimeste erinevat mõtlemist. Suurte andmestike koostamisel kasutati samal eesmärgil ühe ja sama küsimuse näitamist mitmele inimesele.

Küsimuste esitamisel proovisin lähtuda SQUADi puhul annoteerijatele antud juhistest esitada „raskeid“ küsimusi ja kasutada võimalusel küsimuses teisi sõnu kui vastustes. Kasutasin SQUADi esimese, mitte teise versiooni loogikat ehk esitasin ainult selliseid küsimusi, millele oli etteantud lõigus vastus olemas.

Andmestikust kolm lõiku osutusid sellisteks, mille kohta polnud küsimusi võimalik esitada (koosnesid loeteludest, esitasid mitte-entsüklopeedilises stiilis autori vabu mõtteavaldusi või olid kirjutatud arusaamatult).

Kokku kogusin 1115 küsimust ja 1668 vastust (keskmiselt 1,5 erinevat vastust küsimuse kohta, võrdluseks on SQUADi arendus-andmestikus see näitaja 1,8 ehk sarnane).

Töö käigus ilmnes, et küsimuste küsimise protsessis on mitmeid subjektiivseid nüansse. Tõenäoliselt küsiksid eri inimesed erinevaid küsimusi, ja võibolla leiaksid ka selliseid vastuse variante, mida mina ei leidnud. Selles mõttes tagaks paljude inimeste kaasamine universaalsema tulemuse.

Erinevalt inglise keelest, kus käändeid on ainult kaks, kujunes eesti keele puhul ka olukord, kus vastus ei ole sellises käändes, nagu küsimus eeldaks. Võib öelda, et tegu ei olegi otseselt vastusega, vaid tõesti, lühima tekstikatkega, mis vastust sisaldab. See käänete sobimatus võib osutuda probleemiks või vähemasti muudab ülesande natuke teistsuguseks kui ingliskeelse teksti puhul.

Minu kogutud andmestik on suuruselt kindlasti sobilik mudeli testimiseks. Kui aga kasutada osa andmestikust ka treenimiseks, võib osutuda vajalikuks selle täiendamine.

Näidisküsimusi:


```
{
    "context": "Tänapäeval on teadlaste seas kõige soositum nn katastroofihüpotees ehk hiiglasliku kokkupõrke
        hüpotees, mille esitasid kolm teadlast: Bill Hartmann, Roger Phillips ja Jeff Taylor 1980ndatel. Selle
        kohaselt langes Maale üsna tema moodustumise algjärgus ligikaudu Marsi-suurune taevakeha, millele on
        antud nimi Theia. Kokkupõrke tagajärjel eraldus Maast hulgaliselt materjali, millest moodustus Maa
        kaaslane Kuu. Selle plahvatuse energia pani muuhulgas aluse Maa kihilisele ehitusele. Maa sulas ning
        koostiselemendid hakkasid gravitatsiooniliselt diferentseeruma. Sellest ajast on Maal rauast tuum.",
    "qas": [
        {
            "question": "Kes on nn katastroofihüpoteesi autorid?",
            "id": "e949cd45-fd69-4c06-a6f1-275ce32b5779",
            "answers": [
                {
                    "answer_start": 134,
                    "text": "Bill Hartmann, Roger Phillips ja Jeff Taylor"
                }
            ]
        },
        {
            "question": "Kuidas kutsutakse hiiglasliku kokkupõrke hüpoteesi teise nimega?",
            "id": "ce75df93-755d-4ece-818c-dafd51ce90a5",
            "answers": [
                {
                    "answer_start": 47,
                    "text": "katastroofihüpotees"
                },
                {
                    "answer_start": 44,
                    "text": "nn katastroofihüpotees"
                }
            ]
        },
        {
            "question": "Kui suure taevakehaga Maa vastavalt katastroofihüpoteesile kokku põrkas?",
            "id": "a676c3db-f61f-42ef-bd94-891f816bb50d",
            "answers": [
                {
                    "answer_start": 252,
                    "text": "ligikaudu Marsi-suurune"
                },
                {
                    "answer_start": 262,
                    "text": "Marsi-suurune"
                }
            ]
```


5. Töö ülesehitus

Kavatsen ehitada kavandatavad mudelid Huggingface keskkonnas (26) olemasolevate mudelite ja pakettide abil. Tegu on keskkonnaga, mille kaudu on kättesaadav suures koguses keelemudeleid, sh kõik mudelid, mis on olulised minu töö jaoks ja mida mainisin punktis 4.1.

Kavatsen kasutada mudelite treenimiseks Google Colab keskkonda ning kui selle ressurssidest jääb puudu, siis juhendaja kaudu ülikooli poolt pakutavat ligipääsu masinõppeks sobilikele masinatele. Olen treenimisega juba alustanud, aga väga algusjärgus.

6. Tulemuste valideerimine

Tulemuste valideerimiseks kasutatakse küsimus-vastus süsteemidega tegelevates uurimistöödes reeglina kahte mõõdikut – „exact match“ (vaatab, kas süsteemi poolt pakutud vastus vastab täpselt mõnele annoteeritud tõesele vastusele) ja F1-skoor (hindab, kui palju sõnu süsteemi poolt pakutud vastuses kattub annoteeritud tõeste vastustega – võtab aluseks kõige kõrgema skooriga vastusevariandi). Iga küsimus-vastus paari kohta saadud mõõdikute põhjal leitakse kogu testandmestiku aritmeetilised keskmised, mida mudeli töö hindamiseks kasutataksegi. Kavatsen rakendada samu mõõdikuid.

7. Viited

1. **Yse, Diego Lopez.** Your Guide to Natural Language Processing (NLP). *Towards Data Science.* [Võrgumaterjal] 2019. a. https://towardsdatascience.com/your-guide-to-natural-language-processing-nlp-48ea2511f6e1.
2. **Kratzwald, Bernhard ja Feuerriegel, Stefan.** Putting Question-Answering Systems into Practice: Transfer Learning for Efficient Domain Customization. *arXiv.org.* [Võrgumaterjal] 2018. a. https://arxiv.org/abs/1804.07097.
3. **Devlin, Jacob.** Contextual Word Representations with BERT and Other Pre-trained Language Models. *Stanford University.* [Võrgumaterjal] 2020. a. https://web.stanford.edu/class/cs224n/slides/Jacob_Devlin_BERT.pdf.
4. **Miaschi, Alessio ja Dell'Orletta, Felice.** Contextual and Non-Contextual Word Embeddings: an in-depth Linguistic Investigation. *ACL Anthology.* [Võrgumaterjal] 2020. a. https://www.aclweb.org/anthology/2020.repl4nlp-1.15/.
5. **Devlin, Jacob, et al.** BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv.org.* [Võrgumaterjal] 2018. a. https://arxiv.org/abs/1810.04805.
6. **Tanvir, Hasan, Kittask, Claudia ja Sirts, Kairit.** EstBERT: A Pretrained Language-Specific BERT for Estonian. *arXiv.org.* [Võrgumaterjal] 2020. a. https://arxiv.org/abs/2011.04784.
7. **Google Research.** BERT source code in GitHub. *GitHub.* [Võrgumaterjal] 2018. a. https://github.com/google-research/bert.
8. —. Multilingual BERT source code in GitHub. [Võrgumaterjal] 2019. a. https://github.com/google-research/bert/blob/master/multilingual.md.
9. **Chan, Branden.** XLM-RoBERTa: The alternative for non-english NLP. [Võrgumaterjal] 2020. a. https://medium.com/deepset-ai/xlm-roberta-the-multilingual-alternative-for-non-english-nlp-cf0b889ccbbf.
10. **Conneau, Alexis, et al.** Unsupervised Cross-lingual Representation Learning at Scale. *arXiv.org.* [Võrgumaterjal] 2019. a. https://arxiv.org/abs/1911.02116.
11. **Telmo Pires, Eva Schlinger, Dan Garrette.** How multilingual is Multilingual BERT? [Võrgumaterjal] 2019. a. https://arxiv.org/pdf/1906.01502.pdf.
12. **Kittask, Claudia, Milintsevich, Kirill ja Sirts, Kairit.** Evaluating Multilingual BERT for Estonian. *arXiv.org.* [Võrgumaterjal] 2020. a. https://arxiv.org/abs/2010.00454.
13. **Stanford NLP Group.** The Stanford Question Answering Dataset. [Võrgumaterjal] https://rajpurkar.github.io/SQuAD-explorer/.
14. **Rachel Keraron, Guillaume Lancrenon, Mathilde Bras, Fred Eric Allary, Gilles Moyse, Thomas Scialom, Edmundo-Pavel Soriano-Morales, Jacopo Staiano.** Project PIAF: Building a Native French Question-Answering Dataset. [Võrgumaterjal] 2020. a. https://www.aclweb.org/anthology/2020.lrec-1.673.pdf.
15. **Rajpurkar, Pranav, Jia, Robin ja Liang, Percy.** Know What You Don't Know: Unanswerable Questions for SQuAD. *arXiv.org.* [Võrgumaterjal] 2018. a. https://arxiv.org/abs/1806.03822.
16. **Tomasz Jurczyk, Amit Deshmane, Jinho D. Choi.** Analysis of Wikipedia-based Corpora for Question Answering. [Võrgumaterjal] 2018. a. https://arxiv.org/pdf/1801.02073.pdf.
17. **Yang, Yi, Yih, Wen-tau ja Meek, Christopher.** WikiQA: A Challenge Dataset for Open-Domain Question Answering. *ACL Anthology.* [Võrgumaterjal] 2015. a. https://www.aclweb.org/anthology/D15-1237/.
18. **Jiahua Liu, Yankai Lin, Zhiyuan Liu, Maosong Sun.** XQA: A Cross-lingual Open-domain Question Answering Dataset. [Võrgumaterjal] 2019. a. https://www.aclweb.org/anthology/P19-1227.pdf.

19. **Chia-Hsuan Lee, Hung-Yi Lee.** Cross-Lingual Transfer Learning for Question Answering. [Võrgumaterjal] 2019. a. https://arxiv.org/pdf/1907.06042.pdf?.
20. **Pengyuan Liu, Yuning Deng, Chenghao Zhu,Han Hu.** XCMRC: Evaluating Cross-lingual Machine Reading Comprehension. [Võrgumaterjal] 2020. a. https://arxiv.org/pdf/1908.05416.pdf.
21. **Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, Holger Schwenk.** MLQA: Evaluating Cross-lingual Extractive Question Answering. [Võrgumaterjal] 2020. a. https://arxiv.org/pdf/1910.07475.pdf.
22. **Liu, Yinhan, et al.** RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv.org.* [Võrgumaterjal] 2019. a. https://arxiv.org/abs/1907.11692.
23. **Wikimedia.** Wikimedia Downloads. *Wikimedia.* [Võrgumaterjal] 2020. a. https://dumps.wikimedia.org/backup-index.html.
24. **Project Nayuki.** Computing Wikipedia's internal PageRanks. *Project Nayuki.* [Võrgumaterjal] 2016. a. https://www.nayuki.io/page/computing-wikipedias-internal-pageranks.
25. **Mikaelyan, Felix ja Jain, Rahul.** cdqa-suite. *GitHub.* [Võrgumaterjal] 2019. a. https://github.com/cdqa-suite/cdQA-annotator.