TARTU ÜLIKOOL

MATEMAATIKA-INFORMAATIKATEADUSKOND

Arvutiteaduse instituut

Informaatika eriala

Anti Torp

Eestikeelsetest tekstidest akronüümide ja nende vastete ekstraheerimine

Bakalaureusetöö (6 EAP)

Juhendaja: prof. Mare Koit

Autor:..... “.....” juuni 2011

Juhendaja:..... “.....” juuni 2011

Lubada kaitsmisele

Professor..... “.....” juuni 2011

TARTU 2011

Sisukord

1. Sissejuhatus

Akronüümid on lühendid, mis on moodustatud nime või fraasi esitähtedest (nagu USA) või sõnade komponentidest (nagu Benelux: Belgium-Netherland-Luxembourg).

Kui akronüümide tekstist leidmine on triviaalne tegevus nii inimesele kui ka masinale, siis akronüümidele vastete (lahendite) leidmine enam ei ole. Masin võib jääda hätta, sest akronüümide kirjapanekut ja struktuuri mõjutavad näiteks ajaloolised tagamaad ja teised keeled (senimaani pole tehtud ühtegi rahvusvahelist kokkulepet akronüümi moodustamise osas) [2]. Inimene võib jääda hätta, sest kirjatüki autor on pidanud akronüümi enesestmõistetavaks, aga lugeja ei ole selle valdkonna lühenditega tuttav. Võib ka olla, et kirjatüki algul küll mainiti akronüümi tähendust, aga esmakordsel lugemisel ei ole see lugejale meelde jäänud. Lugejapoolset probleemi saab lahendada masinapoolse abiga, kus eelnevalt mainitud lühendid lugejale taaskuvatakse.

Sellisel viisil tekkivat teadmusbaasi saab rakendada programmidele, mis sirvivad tekste. Näiteks saaksid veebibrauserid antud teavet kasutada - hiirega akronüümile liikudes kuvatakse kõige tõenäolisem(ad) vaste(d).

Akronüümi ja selle vaste leidmist võib käsitleda ka anafooride leidmise ülesande osana. Seega saab see töö olla anafooride lahendite leidmise üheks alamülesandeks.

Bakalaureusetöö koosneb järgnevatest osadest: ülevaade akronüümide erinevatest ekstraheerimisviisidest (peatükk 2); sh eestikeelsetest tekstidest akronüümide ekstraheerimise problemaatikast (2.4), eestikeelsetest tekstidest akronüümide ja nende vastete ekstraheerimise programm (peatükk 3, sh tulemuste analüüs ja edasiarendusvõimalused). Viimaks kokkuvõtted eesti ja inglise keeles ning kirjanduse loetelu.

Tööle on lisatud CD peatükis 3 kirjeldatud programmi lähtekoodi, andmete, käivitamisjuhendi ja käesoleva töö tekstiga. Programmi loomisel on kasutatud arenduskeskkonda [37] ja lähtekoodis on kasutatud lõike allikatest [4], [5].

2. Akronüümide ja nende vastete tekstist ekstraheerimine

2.1 Terminoloogia

Järgneva parema mõistmise huvides toon esmalt töös esinevate oluliste mõistete selgitused. Kasutatud on allikaid [1], [2] ja [3].

Akronüüm	- vt. sissejuhatus.
Akronüümi vaste	- akronüümi lahend või lahtikirjutus (*expansion*). Näiteks kui USA on akronüüm, siis selle vaste on *United States of America.*
Vaste pikkus	- Sõnade arv (*length)* vastes ehk lahendis.
Märgis	- ehk *token*. Edaspidi kasutatakse seda mõistet akronüümide vastete genereerimisel ja määramisel, kus märgiseks loetakse lause iga üksikosa, mis on teistest eraldatud tühiku, tabulatsiooni või reavahetusega (*whitespace*).
Saak	- Leitud õigete lahtikirjutuste arvu ja andmetest kõikide lahtikirjutuste arvu suhe (*recall*).
Tugivektormasin	- SVM (*support vector machine*): meetod, mis üritab sisestatud andmetel vahet teha, jagades need kahte klassi (selles töös akronüümide vasteteks ja mitte-vasteteks).
Täpsus	- Leitud õigete akronüümide lahtikirjutuste arvu ja kõikide leitud lahtikirjutuste arvu suhe (*precision*).
Tühemik	- ehk *whitespace*: tühik, tabulatsioon (\t) või reavahetus (\n).
Vahesõne	- Akronüümi ja tema vaste vahel olev suvaline sõne (*offset*). Sõne pikkusühikuks on märgis (*token*).

2.2 Akronüümide ekstraheerimisviisid

On palju viise, kuidas akronüümide vasteid teada saada. Järgnevas on toodud mõned näited [1].

Esmalt, akronüümi vaste võib leida andmebaasist. Leidub mitmeid akronüümide andmebaase (näiteks: www.acronymfinder.com), mida koostatakse kas osaliselt või täielikult käsitsi ja seepärast ei pruugi need väga aja– ja valdkonnakohased olla.

Tüüpilised lähenemised automaatsele ekstraheerimisele on reegli- ja mustripõhised (*rule and pattern based*). Reeglid ja mustrid peavad inimesed ise koostama ja kohandama (*tuning*). Ekstraheeritakse need akronüümid, mis vastavad olemasolevatele reeglitele või mustritele. Sellisel juhul tuvastatakse akronüüm ja selle vaste nö ühe käiguga.

Kolmandaks lähenemiseks (mida on katsetatud inglise keele jaoks) on tugivektormasina (SVM) kasutamine [1]. Esmalt tuvastatakse akronüümid nende omaduste kaudu:

1. akronüüm on kaks kuni kümme sümbolit pikk, millest kõige rohkem üks on tühemik (*whitespace*),
2. selle esimene sümbol on tähestiku täht või number. Vähemalt üks täht on suurtäht (*capital letter*),
3. see ei ole sõnaraamatu sõna, ei ole inimese ega asukoha nimi. Samamoodi ei ole see nö stoppsõnade nimekirjas.

Eeldusel, et akronüümi vaste esineb tekstis alati akronüümi läheduses ja isegi samas lauses, saab iga sõnataoline üksus märgise (*token*). Iga kirjavahemärk saab ühe märgise. Seejärel genereeritakse akronüümi võimalikud vasted, arvestades reegleid:

a) maksimaalne akronüümi ja selle vaste vaheline kaugus ehk *maxoffset* on 10 märgist.
b) akronüümi vaste maksimaalne pikkus (seda mõõdetakse märgistes) peab olema minimaalselt akronüümi enda pikkus + 5 ja maksimaalselt akronüümi kahekordne pikkus.

Järgnevalt üritab tugivektormasin eelneval sammul saadud vastete kandidaatidest õige valida, kasutades eelnevalt treenitud andmetest saadud mudelit. Tugivektormasina mudeli treenimisel arvestatakse akronüümi pikkust, väike- ja suurtähti, kandidaadis eessõnade

asumist, esimesi/viimaseid vaste sõnu ja asjaolu, et akronüümi esitähed on samad, mis selle vaste kandidaadil.

Kuigi eelpool toodud meetodid on üksteisest eristatavad, kombineerivad rakendused tavaliselt erinevate meetodite tulemusi. Peaaegu kõik rakendused kasutavad näiteks stoppsõnade nimekirja. Tugivektormasina rakenduses kasutatakse reegleid (a ja b) vähendamaks võimalike genereeritavate vastete hulka [1].

2.3 Ekstraheerijate tulemuste mõõtmine

Lähtudes asjaolust, et seni loodud ekstraheerijad on valdkonnaspetsiifilised, on erinevate valdkondade põhjal loodud akronüümide ekstraheerijaid problemaatiline võrrelda. Raskus tuleneb sellest, et neid programme luuakse ja testitakse üpriski erinevatel andmestikel. Andmestike suurus sõltub omakorda valdkonnas kasutatavate akronüümide kasutussagedusest ja hulgast.

Siiski on välja kujunenud karakteristikud, mille abil ekstraheerijaid iseloomustatakse – täpsus ja saak. Tabelis 1 on toodud eespool mainitud meetodite mõnede rakenduste karakteristikud.

Meetod	Täpsus (*precision*) %	Saak (*recall*) %	Andmestik
SVM	90,9	84,1	UCI [7]
SVM	89,4	83,4	W3C [1]
Mustripõhine	81,6	82,0	UCI
Mustripõhine	84,9	82,1	W3C
Reeglipõhine (AcroMed [8])	98	72	Medline andmebaas

Tabel 1. Näited erinevate meetodite täpsusest ja saagisest [1], [9].

[1] http://research.microsoft.com/en-us/um/people/nickcr/w3c-summary.html

2.4 Akronüümid ja nende vastete esinemine eestikeelsetes tekstides

Tabelis 2 on toodud mõned näited eestikeelsetes tekstides esineda võivate akronüümide kohta. Näited on leitud, kasutades Tartu Ülikooli tekstikorpuse kasutajaliidest [6] ning tekste [10-36].

Akronüüm	Tüüp, seletus
Eesti Tööandjate Ühenduste Keskliit (ETÜK)	Akronüüm ja selle vaste asuvad teineteise vahetus läheduses . Vasteks oleva fraasi esitähed moodustavad täpse akronüümi.
Keskkonnainspektsioon (KKI)	Akronüüm on moodustatud liitsõnast.
Põhja-Eesti Regionaalhaiglalt (PERH)	Vastes on liitsõna(d).
Eesti Tööstuse ja Tööandjate Keskliit (ETKL)	Vaste on moodustatud nii "loovalt", et sõnade esitähtedest ei moodustu täpne akronüüm.
kroonilises obstruktiivses kopsuhaiguses (KOK) Einsteini teleskoop (ET)	Vaste koosneb täielikult või osaliselt väikese algustähega algavatest sõnadest.
digitaalse infrapuna-ruumiseire projekti SDSS (ingl Sloan Digital Sky Survey)	Eestikeelne akronüümi tõlgendus on ühel pool akronüümi ja ingliskeelne vaste teisel pool.
BRIC (Brazil, Russia, India, China)	Vasteks on loetelu.

Tabel 2. Akronüümide näited.

Väga tuntud akronüümi vastet tavaliselt tekstis ei selgitata (näiteks: EL – Euroopa Liit). Akronüümist endast võib olla saanud oskussõna, mille esialgset vastet pole samuti vaja selgitada (näiteks: NASA ja Laser - *Light Amplification by Stimulated Emission of Radiation*).

Eestikeelsetes tekstides kasutatavate akronüümide omapäraks on tõik, et paljud neist on võetud ja/või tõlgitud ingliskeelsetest tekstidest.

Võõrkeelse akronüümi eesti keelde tõlgitud vastest (*expansion*) koorub probleem - akronüümi vaste esitähed ja akronüüm pole enam kooskõlas. Näiteks: Euroopa Julgeoleku-

ja Koostööorganisatsioon (OSCE), kus OSCE tähendab *Organization for Security and Co-operation in Europe*.

Seega saab eestikeelsetest tekstidest akronüümide ja nende vastete leidmise kohta väita järgnevat:

1. kõikide akronüümide vasteid ei saa ja ei olegi vaja tekstist eraldada.
2. Ülesanne on suures osas sarnane ingliskeelsetest tekstidest ekstraheerimisega.
3. Eestikeelsetest tekstidest võib leida rohkem tõlgitud akronüümide vasteid.
4. Vastel pole sama palju sõnu ja seega ka suuri esitähti, sest eesti keeles kirjutatakse mõned sõnad teisiti - üldiselt kokku või sidekriipsuga (koostöö- organisatsioon on ingl. k. *Co-operation Organization*).
5. Esitähtede järjestus on muutunud (EJK, kus Euroopa on esimesel kohal ja OSCE, kus Euroopa on viimasel kohal).
6. Akronüümi vaste sõnu eraldavad tühikud või sidekriipsud.
7. Tõlgitud sõna esitäht ei pruugi jääda samaks (Julgeolek on ingl. k. *Security*).

3. Eestikeelsetest tekstidest akronüümide ja nende vastete ekstraheerimise programm

Käesolevas bakalaureusetöös loodud programm kasutab akronüümide tuvastamiseks reegleid ja mustreid. Reeglid määravad, millal mingisuguseid mustreid kasutatakse. Eesmärgiks on saavutada ekstraheerija pigem kõrge täpsus (*precision*) kui saak (*recall*).

Järgnevalt tuuakse välja käesolevas töös koostatud akronüüme ja nende vasteid eraldava programmi eesmärgid (nõuete kujul), algoritm, tulemused ja viiakse läbi tulemuste analüüs.

3.1 Ekstraheerija nõuded

Ekstraheerijale kehtestatakse järgmised nõuded. Programm leiab akronüümide vasted kui:

1. akronüümid pikkusega 2-10 on kujul XYZ (pikkus 3) ja vahetult selle ees või järel on akronüüm lahti seletatud kujul XYZ (Xxxx Yyyy Zzzz).
2. Kui need on kujul 2MRS (ingl 2MASS Redshift Survey) ehk akronüümi vahel on nö vahesõne (*offset)*.
3. Kui akronüümi vaste vahel on artiklid, sidesõnad jm.
4. Kui akronüümi lähedal on pikk sõna (ilmselt liitsõna), mis algab sama algustähega. Näiteks "... gammakiirguspurskeks (GRB-ks) …" GRB – Gamma-ray burst või "positronemissioontomograafia (PET)".
5. Kui leiduvad akronüümide vasted, millel ei ole kõiki suuri esitähti, millest akronüüm koosneb ja mis asuvad akronüümile väga lähedal. Näiteks: Eesti Arstiteadusüliõpilaste Selts (EAÜS).
6. Kui tekstis on akronüüm kujul XYZ ja suvaliselt tekstis esineb ka vastav trigramm Xxxxx Yyyyy Zzzzz. Näide: esimeses lauses mainitakse "Eesti Loomakaitse Seltsi" ja järgnevates lausetes hakatakse kasutama akronüümi ELS, siis on põhjust arvata, et esialgne fraas ongi ELS-i vasteks.

Programm ei leia akronüümide vasteid kui:

1. akronüümi vaste tõlkel ei ole sarnasusi akronüümi esitähtedega (vt. eesppool esinenud näide OSCE).
2. Kui see paikneb lauset alustava sõna järel ja lauses esineb miskipärast sama või sarnane akronüüm (näiteks: "… Ajakirjas Astrophysical Journal AAJ …").
3. Kui akronüümiks on mõõtühik (mm, cm, mm/Hg), sest neid on lõplik hulk.
4. Kui vaste koosneb sõnade komponentidest (nagu Benelux: Belgium-Netherland-Luxembourg).

Programmi arendamisel ja testimisel kasutati järgmisi andmeid, millele esitati järgmised nõuded:

1. valiti artiklid, mis ilmusid Delfi Forte rubriigis "Teadus ja Loodus" ajavahemikul 15.-23. mai 2011 ja milles esinesid akronüümid (artikleid, milles need puuduvad, ei loetud andmestikku v.a [8],[12]).
2. Artiklites ei kasutata tavaliste sõnade kirjapanekul läbivalt suurtähti.
3. Iga artikkel on eraldi failis.
4. Artikli iga lause peab olema kirjavahemärkidega eraldatud.

3.2 Algoritm

1. Lugeda tekst(id) faili(de)st mällu.
2. Leida tekstist kõik akronüümid regulaaravaldise '[A-ZÄÖÜÕ0-9][A-ZÄÖÜÕ]{1,10}' järgi ja salvestada.
3. Igast tekstist leitud akronüümidest teha nende vastete oletatavad regulaaravaldised ja otsida vasteid samast tekstist. Kui vaste mingil etapil leitakse, siis mitte edasi otsida.

   Leitakse tüüpjuhud:

   a) akronüümi vaste ja akronüüm on lähestikku ja üks neist on eraldatud sulgudega.
   b) Kui akronüümi vaste asub tekstis suvalisel kohal.
   c) Kui akronüümi vaste vahel asuvad artiklid, eessõnad, siis arvestada nende olemasoluga.

d) Kui akronüümi ees või järel asub pikk liitsõna, mis algab esimese akronüümi algustähega, siis lugeda see vasteks.
e) Kui akronüüm on pikem kui 2 tähte, siis üritada leida suurte ja väikeste algustähtedega algavaid sõnu tekstist.
f) Kui eelnevatest tüüpjuhtudest ei ole vastet leitud ja kui akronüüm on pikem kui 3 tähte (n), siis võtta kombinatsioonid n- elemendist n-1 kaupa ja luua iga kombinatsiooniga regulaaravaldis.

4. Salvestada kõik uued akronüümid ja nende vasted.

3.3 Tulemused

Tabelis 3 on toodud ülevaade artiklites [10-36] esinevate akronüümide ja seal leiduvate vastete kohta. Lisaks on toodud 'Vasteta akronüümid', mille vastet ei esinegi vastavas artiklis. Järjekorranumbrid 1 – 27 tähistavad vastavalt artikleid [10]-[36]

Andmed [11 - 36]	Leitud akronüümide arv	Vasteta akronüümid	Akronüümid ja nende vasted, mida oli võimalik leida
1	3	USA, CNN,	DAIA (Delegación de Asociaciones Israelitas Argentinas)
2	2	3D	2MRS (ingl 2MASS Redshift Survey) 2MASS (ingl Two-Micron All-Sky Survey), Ameerika astronoomialiidu AAS
3	1	Vale: 9B (090429B)	gammakiirguspurskeks (GRB-ks)
4	2		Rahvusvaheline Vähiuuringute Keskus (IARC)
5	1	BBC	
6	2	ERR, NASA	
7	1		UFR (ingl Unified Frame of Reference)
8	3	USA, TTÜ	EL ... Euroopa Liit
9	4		ELI – extreme light infrastructure PALS (ingl Precision Automated Laser Signals)

			seitsmenda sõrestikprogrammi (FP7) Prantsusmaa riiklikust teadusuuringute keskusest CNRS
10	1		Eesti Ornitoloogiaühing (EOÜ)
11	1		SDSS (ingl Sloan Digital Sky Survey)
12	-		
13	2	ERR, AÜ	
14	1	VIDEO	
15	1	OÜ	
16	1	ERR	
17	1		riikliku teadusfondi NSF
18	-		
19	2	BBC	Prantsuse riikliku teadusuuringute keskuse CNRS
20	7	LIGO, LISA, ASPERA, ERR, NASA, TAMA	Einsteini teleskoop (ET)
21	2		ELS … Eesti Loomakaitse Seltsile, ELS … Eesti Loomakaitse Selts
22	1		positronemissioontomograafia (PET)
23	1		TÜ … Tartu Ülikooli
24	1		elektroentsefalograafia (EEG)
25	5	CO, EOKL, MW	EE Eesti Energia, EL Euroopa Liidu
26	3	NASA, USA	MPCV (ingl Multi-Purpose Crew Vehicle)
27	2		Arstiteaduskonna Üliõpilasesindajate Kogu (ATÜK), Eesti Arstiteadusüliõpilaste Selts (EAÜS)

Tabel 3. Leitud akronüümid – kõikide akronüümide arv, mis programm leidis artiklitest [10-36] leidis. Vasteta akronüümid – akronüümid, mis tekstis leidusid, aga vasteid tekstis ei olnud. Akronüümid, mille vasteid oli võimalik leida – vastavas tekstis oli vaste kirjas.

Tabelis 4 on toodud leitud akronüümid ja nendele leitud vasted. Õiged ja samaväärsed vasted on eraldatud leitud valedest vastetest.

Nr	Akronüüm	Õige ja samaväärne vaste	Vale vaste
1	DAIA	Delegacion de Asociaciones Israelitas Argentinas	
2	SDSS	Sloan Digital Sky Survey ingl Sloan Digital Sky Survey	
3	2MRS	2MASS Redshift Survey	2MRS=MASS Redshift Survey
4	ELS	Eesti Loomakaitse Selts Eesti Loomakaitse Seltsile	
5	PET	positronemissioontomograafia	
6	TÜ	Tartu Ülikooli	
7	EEG	elektroentsefalograafia	
8	EL	Euroopa Liidu	Euroopa Leiutaja
9	EE	Eesti Energia	
10	MPCV	Multi-Purpose Crew Vehicle	
11	ATÜK	Arstiteaduskonna Üliõpilasesindajate Kogu	Tartu Ülikooli Kliinikumi
12	EAÜS	Eesti Arstiteadusüliõpilaste Selts	
13	GRB	gammakiirguspurskeks	
14	UFR	ingl Unified Frame of Reference	
15	eli	extreme light infrastructure	
16	PALS	Precision Automated Laser Signals	

Tabel 4. Akronüümid ja nendele leitud vasted.

Tabelis 5 on toodud akronüümid ja nende vasted, mis jäid tekstist leidmata. Lisatud on leidmata jäämise põhjus.

	Akronüümid ja nende vasted tekstis (8)	Põhjus
1	"... Einsteini teleskoop (ET) ..."	Tüüpjuht f (vt. Ptk 3.2) tunneks selle ära siis, kui

		akronüümi pikkus oleks suurem kui 3.
2	"... Eesti Ornitoloogiaühing (EOÜ) ..."	Tüüpjuht f tunneks selle ära siis, kui akronüümi pikkus on suurem kui 3.
3	"... Prantsuse riikliku teadusuuringute keskuse CNRS.."	Akronüümi vaste on tõlgitud
4	"... riikliku teadusfondi NSF ..."	Akronüümi vaste on tõlgitud
5	"... seitsmenda sõrestikprogrammi (FP7) ..."	Akronüümi vaste on tõlgitud
6	"... Rahvusvaheline Vähiuuringute Keskus (IARC) ..."	Akronüümi vaste on tõlgitud
7	"... Ameerika astronoomialiidu AAS."	Kui tüüpjuhud d ja f oleks kombineeritud
8	"... 2MASS (ingl Two-Micron All-Sky Survey) ..."	'2' ei ole sama mis 'Two'

Tabel 5. Leidmata vasted.

Tabelis 6 on toodud eestikeelsetest tekstidest akronüümide ekstraheerija tulemuste karakteristikud ja selle väärtus.

Karakteristik	Väärtus
Leitud akronüümide arv	43
Akronüümide õigete vastete arv, mille program leidis	16
Akronüümide arv, mille vaste on tekstis olemas	24
Leidmata vastete arv	8
Valede akronüümide vastete arv	3
Täpsus (*Precision*)	16/19 = 84,2%
Saak (*Recall*)	16/24 = 66,6%

Tabel 6. Ekstraheerija tulemused.

Arvestada tuleb, et testandmetes esinevad ainult mõned Eesti kirjakeele korpusest [6] autorile silma hakanud tüüpjuhud ja kindlasti leidub mõni tüüpjuht, mis on tähelepanuta jäänud.

Kuna testandmed on suhtelised lühikesed artiklid, siis on mõistlik otsida akronüümide vasteid kogu artikli ulatuses (vt. tüüpjuht b). Suuremamahulise teksti puhul tuleks säärase üldistuse puhul ettevaatlik olla.

3.4 Edasiarendusvõimalused

Programmi saab edasi arendada vähemalt järgmistel viisidel:

1. Kui antud ekstraheerijat kasutada mõnel uudiste veebilehel ja tavakasutajale kuvada võimalike akronüümide vasteid ja lasta valida, milline on õigem, siis saaks vastava akronüümi ja selle vaste vahelist sidet "tugevdada".
2. Rakendades eesti keele morfoloogilist analüsaatorit leitud akronüümi vastele, peaks teada saama võimalike liitsõnade sõnapiire. Nii saaks väita veelgi kindlamalt, et näiteks akronüümi PET vaste on 'positronemissioontomograafia' ja TALO vaste 'Teenistujate Ametiliitude Organisatsioon '.
3. Rakendada võimalikele akronüümi vastetele ja akronüümi ümbrusele keeletuvastajat, saamaks teada, millises keeles on vaste ja akronüümi ümbrus.
4. Mõõta eelnevalt saadud vaste usaldusväärsust, parandamaks valede akronüümide seast õigete valimise täpsust. Näiteks kui akronüümile ATÜK on leitud vasted 'Arstiteaduskonna Üliõpilasesindajate Kogu' ja 'Tartu Ülikooli Kliinikumi', siis valida nende hulgast õige.

Veel ühe võimaliku edasiarendusena võiks tulevikus proovida tugivektormasina implementeerimist eestikeelsete tekstide põhjal.

4. Kokkuvõte

Töös toodi kirjanduse põhjal välja mitu viisi selle kohta, kuidas eelnevalt on üritatud lahendada akronüümide vastete leidmise probleemi: käsitsi koostatud andmebaasid, reegli- ja mustripõhised lähenemised ja tugivektormasina kasutamine. Selgitati erinevaid ekstraheerimijaid võrdlevaid karakteristikuid ja toodi välja nendega seotud probleemid. Kirjeldati probleeme, mis tekivad eestikeelsetest tekstidest akronüümide vastete ekstraheerimisel.

Töös loodi eestikeelsetest tekstidest akronüümide ja nende vastete ekstraheerija prototüüp, esitati selle eesmärgid, kastutatud algoritm ja programmi testimise tulemused. Põhilised akronüümide ja nende vastete mallid on saadud andmete põhjal, mille seas leidus nii ainult eestikeelseid kui ka tõlgitud tekste (üldiselt olid tekstid tõlgitud inglise keelest ja sisaldasid kohati ingliskeelseid sõnu). Võib ütelda, et kuigi mallid koostati näitepõhiselt, siis vähemasti saadi malle mitme tüüpjuhu kohta.

Prototüüp saavutas täpsuseks (*precision*) 84,2% ja saagiks (*recall*) 66,6%. Need karakteristikud ei ole päris usaldusväärsed, sest suurema ja juhuslikuma andmevalimi korral ei ole alust arvata, et näitajad ikka sama kõrgeks jäävad.

Töös on toodud ka programme edasiarendusvõimalused.

5. Acronym extraction from texts written in Estonian

Summary

The aim of this paper was to give an overview of acronym extraction in general and to try to implement the knowledge on texts written in Estonian. As there is no universal agreement on the definition, it is a vague term. Acronym is an abbreviation formed from the initial components in a phrase [2]. Because of that they can be following: USA meaning 'United States of America' and Benelux meaning 'Belgium-Netherland-Luxembourg'. Here we identify that there are acronyms and their expansions – 'United States of America' would be an expansion for USA.

The two named acronyms are well known and searching for their expansions is unnecessary, however there are more specific acronyms that one can find while reading long scientific texts. In that case, it would be helpful to get an instantaneous recall of possible acronym expansion candidates.

The simplest way to get expansion candidate is to search manually compiled databases. That solution is followed by automated extraction solutions: pattern and rule-based

The general solution for automated acronym extraction is to identify the acronyms and recognize their expansions from surrounding text. This problem gets more difficult when dealing with text written in another language (here we try to solve the problem with Estonian language). The increased difficulty is caused by the fact that a lot of texts are translated from English and some of the acronym expansions are translated, while the acronyms are not. The problem gets worse since Estonian translation of a regular English acronym might be a compound noun. Luckily, all the cases are not so extreme and most acronyms are closely preceded or followed by their expansions.

There are two metrics that are used to describe acronym extractors – precision and recall. Precision measures how many correct expansions are extracted compared to all expansions found. Recall measures how many expansions were identified compared to what was possible to identify.

Lastly, there is an attempt to create prototype extractor for Estonian language using simple regular expressions to match and extract acronyms and their expansions from texts written in Estonian. This attempt is tested on about 30 small articles that contain acronyms.

While the main idea was to get the prototype to match expansions without making too many mistakes, the patterns that were compiled are intended to have as high precision as possible (the prototype scored 84.2%) and leaving questionable expansions out. That is the reason the prototype's recall score was 66.6% (compared to SVM's, which was 84.1%/83.4%).

6. Kirjandus

1. Jun Xu, Yalou Huang. "Using SVM to extract acronyms from text", Soft Computing –SOCO , vol. 11, no. 4, pp. 369-373, 2007 DOI 10.1007/s00500-006-0091-5
2. Wikipedia artikkel "*Acronym and initialism*", http://en.wikipedia.org/wiki/Acronym_and_initialism, (viimati külastatud 24.05.2011).
3. Wikipedia artikkel "*Support vector machine*", http://en.wikipedia.org/wiki/Support_vector_machine, (viimati külastatud 5.06.2011).
4. Roedy Green, Canadian Mind Products. "Regex: Java Glossary", http://mindprod.com/jgloss/regex.html, (viimati külastatud 5.06.2011).
5. Michael Gilleland. "Combination Generator", http://www.merriampark.com/comb.htm, (viimati külastatud 4.06.2011).
6. http://www.cl.ut.ee/korpused/kasutajaliides/, (viimatu külastatud 5.06.211)
7. Hettich S, Bay SD (1999) The UCI KDD Archive. [http://kdd.ics.uci.edu]. Department of Information and Computer Science,University of California, Irvine
8. Pustejovsky J, Castano J, Cochran B, KoteckiM, Morrell M (2001) Automatic extraction of acronym-meaning pairs from MEDLINE databases. Medinfo 10(Pt 1):371–375
9. Ménard, P. A., & Ratté, S. (2010). Classifier-based acronym extraction for business documents. *Knowledge and Information Systems*. Retrieved from http://www.springerlink.com/index/10.1007/s10115-010-0341-9, (viimati külastatud 12.06.2011).
10. Artikkel 1: http://forte.delfi.ee/news/digi/argentina-googleit-sunniti-juudivaenulikke-linke-blokeerima.d?id=46462183, (viimati külastatud 5.06.2011).
11. Artikkel 2: http://forte.delfi.ee/news/teadus/3d-kosmos-valmis-universumi-maa-lahipiirkonna-taiuslikem-ruumiline-kaart.d?id=46625382, (viimati külastatud 5.06.2011).
12. Artikkel 3 : http://forte.delfi.ee/news/teadus/kolmandik-koigist-seni-avastanud-planeetidest-tiirlevad-umber-oma-tahe-mitmekesi.d?id=46624698, (viimati külastatud 5.06.2011).

13. Artikkel 4: http://teadus.err.ee/artikkel?id=4425&cat=1, (viimati külastatud 5.06.2011).
14. Artikkel 5: http://forte.delfi.ee/news/teadus/teadlaste-loodud-robotid-arendavad-uut-omavahelist-suhtluskeelt.d?id=46625034, (viimati külastatud 5.06.2011).
15. Artikkel 6: http://forte.delfi.ee/news/teadus/kolmandik-koigist-seni-avastanud-planeetidest-tiirlevad-umber-oma-tahe-mitmekesi.d?id=46624698, (viimati külastatud 5.06.2011).
16. Artikkel 7: http://forte.delfi.ee/news/teadus/enamik-teadusele-tuntud-dinosaurustest-osutuvad-puhtalt-valjamoeldisteks.d?id=46275481, (viimati külastatud 5.06.2011).
17. Artikkel 8: http://forte.delfi.ee/news/teadus/budapestis-selgusid-euroopa-leiutaja-2011-voitjad.d?id=46275155, (viimati külastatud 5.06.2011).
18. Artikkel 9: http://forte.delfi.ee/news/teadus/euroopa-esimene-superlaser-rajatakse-praha-lahistele.d?id=46272173, (viimati külastatud 5.06.2011).
19. Artikkel 10: http://forte.delfi.ee/news/teadus/anna-teada-kas-oled-sellel-kevadel-juba-linde-vaatlemas-kainud.d?id=46270941, (viimati külastatud 5.06.2011).
20. Artikkel 11: http://forte.delfi.ee/news/teadus/elukolbulikke-planeete-voib-olla-seniarvatust-rohkem.d?id=30444695, (viimati külastatud 5.06.2011).
21. Artikkel 12: http://forte.delfi.ee/news/teadus/ulekaalulisi-eesti-mehi-ohustab-aina-enam-viljatus.d?id=46261207, (viimati külastatud 5.06.2011).
22. Artikkel 13: http://forte.delfi.ee/news/teadus/kodutud-planeedid-voivad-linnuteele-iseloomulikud-olla.d?id=46260877, (viimati külastatud 5.06.2011).
23. Artikkel 14: http://forte.delfi.ee/news/teadus/video-balti-merevaigust-leiti-49-miljonit-aastat-vana-suur-amblik.d?id=46253395, (viimati külastatud 5.06.2011).
24. Artikkel 15: http://forte.delfi.ee/news/teadus/puhapaeval-hakatakse-koerte-elukvaliteeti-parandama.d?id=46245861, (viimati külastatud 5.06.2011).
25. Artikkel 16: http://forte.delfi.ee/news/teadus/meeste-menopaus-on-tavaline-nahtus.d?id=46303717, (viimati külastatud 5.06.2011).
26. Artikkel 17: http://forte.delfi.ee/news/teadus/fukushima-radioaktiivne-moju-ookeanidele-on-tsernobolist-suurusjargu-vorra-korgem.d?id=46303889, (viimati külastatud 5.06.2011).
27. Artikkel 18: http://forte.delfi.ee/news/teadus/ajalugu-aatomituuma-avastamine-ja-tuumafuusika-sai-100-aastaseks.d?id=46311037, (viimati külastatud 5.06.2011).

28. Artikkel 19: http://forte.delfi.ee/news/teadus/amazonases-elab-suguharu-mil-puudub-arusaam-ajast.d?id=46317871, (viimati külastatud 5.06.2011).
29. Artikkel 20: http://teadus.err.ee/artikkel?cat=1&id=4427, (viimati külastatud 5.06.2011).
30. Artikkel 21: http://forte.delfi.ee/news/teadus/ukskoikne-suhtumine-asendamatusse-valvesusteemi-ehk-ketikoerte-piinarikas-elu.d?id=46456035, (viimati külastatud 5.06.2011).
31. Artikkel 22: http://forte.delfi.ee/news/teadus/miks-muudab-soomine-anorektiku-rahutuks.d?id=46474581, (viimati külastatud 5.06.2011).
32. Artikkel 23: http://forte.delfi.ee/news/teadus/vaarikate-ulikoolis-lopeb-esimene-oppeaasta.d?id=46477407, (viimati külastatud 5.06.2011).
33. Artikkel 24: http://forte.delfi.ee/news/teadus/tunned-valu-aja-kaed-risti-ajad-aju-segadusse-ja-hairid-valuaistingu-tekkimist.d?id=46551471, (viimati külastatud 5.06.2011).
34. Artikkel 25: http://forte.delfi.ee/news/teadus/omanike-keskliit-eestisse-tuleb-ehitada-tuumajaam.d?id=46575364, (viimati külastatud 5.06.2011).
35. Artikkel 26: http://forte.delfi.ee/news/teadus/nasa-kavandatav-kapsel-viib-astronaudid-rekordkaugustesse-avakosmosesse.d?id=46581562, (viimati külastatud 5.06.2011).
36. Artikkel 27: http://forte.delfi.ee/news/teadus/taskus-naeb-arstitudengite-salajasi-motteid.d?id=46602244, (viimati külastatud 5.06.2011).
37. Eclipse allalaadimise lehekülg, http://www.eclipse.org/downloads/, Eclipse IDE for Java Developers (92 MB), (viimati külastatud 25.05.2011).

7. Lisa

1. CD-plaat lähtekoodi, andmete, käivitamisjuhendi ja käesoleva töö tekstiga.