STATISTINIŲ LIETUVIŲ KALBOS MODELIŲ KŪRIMAS IR PIRMINIS TYRIMAS

Airenas Vaičiūnas, Gailius Raškinis

Informatikos fakultetas, Vytauto Didžiojo Universitetas, Vileikos g. 8, LT-3035 Kaunas, Lietuva
E-mail: airenas@mail.lt, idgara@vdu.lt

Šiame straipsnyje pristatomas mūsų tyrimas, kurio tikslas sukurti statistinį lietuvių kalbos modelį, tinkamą naudoti šnekos atpažinimo sistemose. Tai pradinis mūsų tyrimas, kuris taip pat yra vienintelis ir pirmasis bandymas sukurti tokio tipo modelį lietuvių kalbai. Straipsnyje aprašyti eksperimentai su trimis mūsų sukurtais statistiniais kalbos modeliais: įprastu 3-gramos modeliu, ir dviem sudėtiniais modeliais, sudaytais iš keleto n-gramų, kurie gauti skaidant žodžius į dvi – prasminę ir morfologinę – dalis. Sudėtiniai modeliai tarpusavyje skiriaisi skaidymo principu (funkcija). Pirmuoju atveju žodis skaidomas pagal galūnių sąrašą, kaip simbolių eilutė į dvi dalis. Antruoju atveju skaidymui panaudotas morfologinis analizatorius, kuris išskaido žodį į jo pagrindinę formą ir gramatinį apibūdinimą. Visais trimis atvejais įvertinta kalbos modelių maišatis[3] ir neaprėpto žodyno dydis[4].

Raktiniai žodžiai: statistinis kalbos modeliavimas, maišatis, n-gramos, kaitomos kalbos.

1. Įžanga

Šnekos atpažinimo sistemos tikslumas gali būti pagerintas panaudojant joje kalbos modelį (toliau KM). Kalbos modelio užduotis – priskirti kiekvienai atpažinimo sistemos generuojamai žodžių sekai tam tikrą "gerumo" įvertį. Dažniausiai tam naudojami statistiniai KM.

Statistinis kalbos modelis - tai modelis, kuris kiekvienai žodžių sekai w_1, … w_N, kur N yra žodžių sekos ilgis, priskiria tikimybę:

$$P(w_1...w_N) = P(w_1)P(w_2 \mid w_1)P(w_3 \mid w_1w_2)...P(w_n \mid w_1...w_{N-1}) = \prod_{i=1}^{N} P(w_i \mid w_1...w_{i-1}), \qquad (1)$$

Praktikoje naudojami paprastesni n-gramos tipo modeliai, t.y. formulė (1) supaprastinama, darant prielaidą, kad i-tasis žodis w_i, priklauso tik nuo n-1 prieš jį einančio žodžio. Labiausiai paplitę trigramos tipo modeliai, kuomet n = 3.

$$P(w_i \mid w_1...w_{i-1}) \approx P(w_i \mid w_{i-n+1}...w_{i-2}w_{i-1}) \qquad (2)$$

Statistinio KM gerumui įvertinti naudojamas *maišytumo* įvertis *PP*. Sakykime kad stebime žodžių seką w_1, … w_{i-1}. Maišytumas parodo, koks vidutinis vienodai tikėtinų žodžių aibės dydis, iš kurios turėtume spėti žodį w_i. Maišytumas tiesiogai susijęs žodžių sekos entropijos įverčiu $\hat{H}$.

$$PP = 2^{\hat{H}}, \qquad (3)$$

$$\hat{H} = -\frac{1}{N}\log_2 P(w_1...w_N). \qquad (4)$$

Dar viena svarbi KM charakteristika - tai *neaprėpto žodyno dydis* (toliau NŽ), kuris parodo kokią nežinomo teksto žodžių dalį sugeba padengti KM žodynas. Svarbus ir KM užimamos *atminties dydis*, kuris net ir šiuolaikiniams kompiuteriams gali būti problematiškas.

2. Panašūs darbai

Pasaulyje daug dirbama kalbos atpažinimo srityje. Neišvengiamai daug dėmesio skiriama ir kalbos modeliams. Anglų kalbai neblogi rezultatai pasiekiami su paprasčiausia trigrama. Deja, paprasčiausios n-gramos neefektyvios labai kaitomos kalboms, tarp jų ir lietuvių kalbai. Kaitomų kalbų atveju, KM žodynas tampa labai didelis ir dėl to suprastėja n-gramos įverčiai [1]. Šią problemą galima spręsti keliais būdais.

---

[3] Perplexity – angl.
[4] Out-off vocabulary – angl.

D.Carter ir bendraautoriai [1] švedų kalbos atpažinimo sistemoje sprendė sudurtinių daiktavardžių problemą. Skaidydami žodžius, jie gavo 2% geresnį NŽ parametrą - nuo 5% sumažino iki 3%. Dėl to jų atpažinimo sistema, į kurią toks KM buvo įjungtas, davė geresnius rezultatus.

C.Martins ir kiti [4] atliko eksperimentus su portugalų kalbos modeliu. Jų kalbos modelyje žodžiai buvo mofologiškai skaidomi į šaknis ir į priesagas, o pats trigramos modelis pakeistas 4-grama.

Prieš skaidymą: P(bem|eles cantam).

Po skaidymo: P(bem|else can tam).

Palyginę rezultatus, gautus naujuoju būdu, su standartine 3-grama, autoriai nustatė, kad naujasis KM turi 29% mažesnį žodyną ir 63% mažesnį KM maišytumą.

Naudodamasis apie 100 mln. žodžių dydžio anglų ir rusų kalbų tekstynais E.Whittaker [5] atliko daug eksperimentų, kurdamas ir lygindamas abiejų kalbų statistinius modelius. Pirmųjų eksperimentų metu jis sukonstravo abiejų kalbų 3-gramos tipo KM. Vėlesniuose – panaudojo žodžių skaidymo į sudėtines dalis principą ir 3-gramą pakeitė 6-grama. Pažymėtina tai, kad jo naudota žodžių skaidymo funkcija nėra fiksuota. Dalyje jo eksperimentų bandoma geriausią žodžių skaldymo būdą parinkti automatiškai.

Tame pačiame darbe autorius aprašė rezultatus, gautus grupuojant žodžius į klases. Jis atliko dviejų tipų grupavimo į klases bandymus:

- $P(w_n \mid w_1 ... w_{n-1}) \approx P_0(w_n \mid C(w_n)) P_1(C(w_n \mid C(w_{n-N+1}) ... C(w_{n-1}))$, (5)

kur C – funkcija priskirianti žodį w klasei $C(w)$: $C : w \to C(w)$, P_1 - klasių 3-grama, P_0 - tikimybinis modelis, kad žodis w priklauso klasei $C(w)$.

- Kitas būdas tiesiogiai prognozuoti žodį pagal prieš jį einančias klases:

$$P(w_n \mid w_1 ... w_{n-1}) \approx P\ (w_n \mid C_{N-1}(w_{n-N+1}) ... C_1(w_{n-1})) \quad (6)$$

E.Whittaker įvertino visų nagrinėtų KM maišatį, NŽ dydį. Tiesiškai sujungdamas tris kalbos modelius, gavo 19% mažesnį maišaties įvertį rusų kalbai ir 11.2% anglų kalbai, lyginant su atitinkamais 3-gramų įverčiais. Tiesinis modelių sujungimas nusakomas bendra (7) formule:

$$P(w_n \mid w_1 ... w_{n-1}) = \sum_{i=1}^{M} \lambda_i P_i(w_n \mid w_1 ... w_{n-1}), \quad (7)$$

kur, P_i - žodžiųsekos tikimybės įvertis naudojant modelį i, M – sujungiamų modelių skaičius.

Nors aukščiau aprašyto tipo statistiniai modeliai jau daugelį metu nagrinėjami įvairioms kalboms, analogiškų darbų atliktų su lietuvių kalba iki šiol nebuvo. Dėl to ne tik įdomu, bet ir būtina sukurti ir išnagrinėti įvairius statistinius lietuvių KM, palyginti jų parametrus su kitų kalbų statistinais modeliais. Tai, kad lietuvių kalbos žodžiai yra labai kaitomi ir sakiniuose nėra griežtos žodžių tvarkos, daro statistinį modelį sudėtingesniu. Taigi, šio darbo tikslas yra įvertinti statistinio kalbos modelio lietuvių kalbai parametrus, išnagrinėti kalbos modelius pagristus žodžių skaidymu į dalis.

3. Lietuvių kalbos statistiniai modeliai

3.1. Modeliavimo duomenys ir įrankiai

Eksperimentai buvo atlikti naudojantis VDU tekstynu[5]. Jame sukaupti įvairaus pobūdžio: moksliniai, grožiniai, publicistiniai ir kitokie. tekstai. Buvo naudojama tekstyno versija turinti 84,202,576 žodžius. Tekstynas buvo padalintas į tris imtis: 98% skirti modelio mokymui, 1% - modelio parametrų optimizavimui 1% - modelio testavimui. Kadangi tekstynas apima didelę žanrų įvairovę, kiekvieno žanro tekstų rinkinys taip pat buvo dalinamas stengiantis išlaikytos šias proporcijas. Išankstinio tekstų apdorojimo metu visi skyrybos ženklai buvo pašalinti ir visi skaičiai pakeisti viena žyme <num>.

Modeliai buvo konstruojami naudojantis CMU-SLMT programinės įrangos paketu[2]. Šis paketas buvo išplėstas ir patobulintas, kad būtų galima dirbti su didesniu nei 65000 žodžių žodynu. Žodžių morfologinė analizė buvo atliekama su programine įranga Lemuoklis[6].

---

[5] Tekstyno pagrindinė kūrėja - habil dr. R. Marcinkevičienė. Tekstynas surinktas ir paruoštas Vytauto Didžiojo Universiteto Kompiuterinės Lingvistikos Centre. http://www.donelaitis.lt/.

3.2. Eksperimentai

Darbo metu buvo atlikti 3 pagrindiniai eksperimentai ir sukurti 3 :statistiniai lietuvių kalbos modeliai: įprastas 3-gramos modelis ir du sudėtiniai modeliai, sudayti iš keleto n-gramų. Abu sudėtiniai modeliai remiasi žodžių skaidymo į prasminę ir morfologinę dalis principu, tačiau naudoja skirtingą skaidymo būdą (funkciją).

3.2.1. Lietuvių kalbos 3-grama

Darbe sukurta standartinė lietuvių kalbos 3-grama. Šiuo atveju žodžio tikimybė skaičiuojama remiantis (2) formule. 3-gramai taikytas Good-Turing glotninimo metodas, bei grįžimo atgal[6] principas[3]. Nustatyti lietuvių kalbos 3-gramos maišaties ir NŽ dydžio įverčiai.

3.2.2. Sudėtiniai statistiniai lietuvių kalbos modeliai

Darbe pasiūlyti nauji sudėtiniai statistiniai lietuvių kalbos modeliai. Jie paremti žodžio skaidymu į dvi dalis, bandant atskirti žodyje esančią prasminę ir morfologinę informaciją. Sakykime, kad turime funkciją *S*, kuri kažkokiu būdu išskaido žodyje esančią informaciją į dvi dalis:

$$S : w \to S(w) = m, e \tag{8}$$

kur *m*- pirmoji žodžio informacijos dalis, o e – antroji. Vienintelis reikalvimas, kuris keliamas tokiai funkcijai - tai atvirkštinės funkcijos S^{-1} egzistavimas, tokios atvirkštinės funkcijos, kuri vienareikšmiškai atstaytų žodį *w* pagal jo sudedamąsias dalis *m* ir *e*.

Pertvarkydami (2) formulę turime:

$$\begin{aligned} P(w_n \mid w_1 ... w_{n-1}) &= P(m_n e_n \mid m_{n-2} e_{n-2} m_{n-1} e_{n-1}) \\ &= P(m_n \mid m_{n-2} e_{n-2} m_{n-1} e_{n-1}) P(e_n \mid m_{n-2} e_{n-2} m_{n-1} e_{n-1} m_n) \end{aligned} \tag{9}$$

Įsivaizduodami, jog pirmoji žodžio informacijos dalis yra tikrai prasminė, o antroji - morfologinė, galime bandyti supaprastinti (9) išraišką. Galime padaryti dvi prielaidas:

- žodžio prasminė informacija nepriklauso nuo anksčiau einančių žodžių morfologinės informacijos;
- žodžio morfologinė informacija nepriklauso nuo anksčiau einančių žodžių prasminės informacijos, bet priklauso nuo to paties žodžio prasminės informacijos dalies.

Šios abi prielaidos leidžia supaprastinti (9) išraišką:

$$P(w_n \mid w_1 ... w_{n-1}) = P(m_n e_n \mid m_{n-2} e_{n-2} m_{n-1} e_{n-1}) \approx P(m_n \mid m_{n-2} m_{n-1}) P(e_n \mid e_{n-2} e_{n-1} m_n) \tag{10}$$

Praktiškai, šiame darbe (10) išraišką dar labiau prastinome, jos narį $P(e_n \mid e_{n-2} e_{n-1} m_n)$ pakeisdami kitų dviejų tikimybių tiesine kombinacija $\lambda P(e_n \mid e_{n-2} e_{n-1}) + (1-\lambda) P(e_n \mid m_n)$. Iš čia:

$$P(w_n \mid w_1 ... w_{n-1}) \approx P_m(m_n \mid m_{n-2} m_{n-1})(\lambda P_e(e_n \mid e_{n-2} e_{n-1}) + (1-\lambda) P_{me}(e_n \mid m_n)), \tag{11}$$

kur $0 \le \lambda \le 1$. Gautasis statistinis kalbos modelis apima tris tikimybinius modelius: prasminių žodžio dalių 3-gramą P_m, morfologinių žodžio dalių 3-gramą P_e ir modelį, kuris pateikia sąlyginę morfologinės žodžio dalies tikimybę, kai žinoma to paties žodžio prasminė informacija P_{me}. Kiekvienas iš šių statistinių modelių lengviau apskaičiuojamas ir reikalauja mažiau atminties nei ekvivalentus bendras žodžių 3-gramos modelis. λ reikšmė buvo surandama eksperimentiškai, parenkant tokią reikšmę, kuri pasiekia geriausią maišytumo parametrą su mokymo optimizavimui skirta duomenų imtimi.

Kaip jau buvo minėta, ištyrėme dvi skirtingas žodžio skaidymo funkcijas *S.*

1. Pirmoji mūsų sukonstruota funkcija bandė atskirti prasminę ir morfologinę žodžio informaciją, skaidydama žodį į dvi dalis kaip simbolių eilutę. Šis skaidymas paremtas intuityviu suvokimu, jog prasminę informaciją saugo labiau pirmoji, o morfologinę – antroji žodžio pusė.

Pirmiausia rankiniu būdu buvo sudatyta galima lietuvių kalbos žodžių galūnių aibė *E*. Į ją surinktos lietuvių kalbos žodžiuose pasitaikančios galūnės: a, ą, ai, ais, aisi, aisiais, …. Skaidymo funkcija *S* patikrindavo, ar nagrinėjamas žodis pasibaigia viena iš *E* aibės galūnių ir, jei rasdavo atitikimą, atitinkamai padalindavo žodį į dvi dalis. Jei ne – visi žodžio simboliai būdavo interpretuojami kaip prasminė informacija *m*, o morfologinei informacijai *e* būdavo priskiriama žymė <empty> (tuščia galūnė) (žr 1 lentelę). Kai kuriuose eksperimentuose į galūnių sąrašą buvo įtraukiami ir 36 prielinksniai, kurie reikalauja konkretaus linksnio ir tuo neabejotinai įtakoja po to einančio žodžio galūnę. Tokiu atveju funkcija *S* visą prielinksnį įrašydavo ir į *m*, ir į *e* žodžio dalis. Eksperimentai atlikti su įvairiomis galūnių aibėmis *E*: nuo 3 iki 489 galūnių.

---

[6] back off – (angl.)

1 lentelė. Žodžio-simbolių eilutės skaidymo į prasminę ir morfologinę dalis pavyzdžiai

Žodis, *w*	**Prasminė informacija, *m***	**Morfologinė informacija, *e***
mes	mes	<empty>
bėga	bėg	a
bėgs	bėg	s
prie	prie	>prie

2. Antroji mūsų sukonstruota funkcija bandė atskirti prasminę ir morfologinę žodžio informaciją, panaudodama morfologinės analizės įrankį Lemuoklis. Lemuokliui pateikus žodį, jis grąžindavo informaciją apie to žodžio pagrindinę formą (pvz. daiktavardžiams – tai vienaskaitos vardininkas), ir jo morfologinę informaciją : kalbos dalis, giminė, skaičius, linksnis, laikas, asmuo ir kt. (žr. 2 lentelę) Kai kuriems daugiaprasmiškiems žodžiams Lemuoklis grąžindavo ne vieną, o kelis gramatinius apibūdinimus. Visų šių apibūdinimų rinkinys šiame darbe buvo traktuojamas kaip dar vienas papildomas morfologinis apibūdinimas. Jei nagrinėjamo žodžio nebūdavo morfologinio analizatoriaus žodyne, žodžio pagrindinė forma būdavo sutapatinama su nagrinėjamu žodžiu, o į gramatinę kategoriją įrašoma <le_ne> (žodis neišanalizuotas).

2 lentelė. Žodžio skaidymao į pagrindinę formą ir gramatines kategorijas pavyzdžiai

Žodis, *w*	**Prasminė informacija, *m***	**Morfologinė informacija, *e***
mes	aš	4000000000210;
medis	medis	1000000002110;
medžiai	medis	1000000002210;1000000002270;
žūb	žūb	<le_ne>

3.3. Rezultatai

3.3.1. Lietuvių kalbos 3-gramos modelio maišytumas ir NŽ

3 lentelė. 3-gramos maišytumas ir NŽ dydis. Kairiajame stulpelyje pateiktas 3-gramos maišytumas, atmetus žodžių poras ir trejetus tekste pasitaikančias tik vieną kartą. Dešinėje pusėje pateiktas 3-gramos maišytumas ir NŽ dydis, neatmetus retų žodžių kombinacijų

Žodžių kiekis	**Atmetus**	**Neatmetus**	
	Maišytumas	**Maišytumas**	**NŽ %**
65000	638.83	414.3	10.92
100000	734.09	449.76	8.25
200000	899.51	512.75	5.05
400000	1052.68	570.32	3.07
500000	1099.27	588.03	2.58
800000	1187.87	621.69	1.82
1000000	1217.19	631.09	1.62
1422746(visi)		-	1.21

3 lentelėje matomi rezultatai, kaip maišytumas ir NŽ kinta, didėjant žodyno kiekiui. Kaip parodė rezultatai, kas buvo netikėta, kad net su visais mokymo imtyje esančiais žodžiais (1.4 mln.) nepavyko pasiekti 1% NŽ rezultato. Anglų kalboje tai pasiekiama su 65000, o rusų su 430000 žodynu[5].

3 lentelėje taip pat matome, kad maišytumas atmetus žodžių poras ir trejetus pasitaikančius tekste tik 1 kartą žymiai padidėjo. Tai reiškia, kad retos žodžių kombinacijos talpina svarbią informaciją, ir be jų maišatis pablogėja nuo 36% 65000 žodynui iki 48% 1 mln. žodynui.

Testiniame tekste buvo apie 15% trigramų, kurios mokymo metu pasitaikė tik vieną kartą. Analogiškuose įvertinimuose rusų ir anglų kalboms šis procentas siekia 7%[5]. Iš to galime spręsti, kad mūsų naudotame tekstyne sukaupta įvairesnio pobūdžio informacija ir 3-gramos įverčiai nėra pakankamai tikslūs.

Lyginant lietuvių kalbos 3-gramatikos maišytumą (5 lentelė) ir NŽ su rusų ir anglų kalbomis [5], pastebime, kad lietuvių kalboje šie parametrai yra didesni. Tai gali būti paaiškinama tuo, kad E.Whittaker maišytumo ir NŽ parametrus vertino naudodamas panašų tekstą. Tuo tarpu mūsų tekstynas parinktas taip, kad teksto pobūdis būtų kuo įvairesnis.

4 lentelė. Rasta dalis trigramų kai vienetinės trigramos buvo atmestos arba paliktos

Žodžių kiekis	Rasta dalis trigramų %, kai vienetinės trigramos nebuvo atmestos	Rasta dalis trigramų %, kai vienetinės[7] trigramos buvo atmestos
65000	55.61	40.07
100000	52.71	36.72
200000	49.33	33.00
400000	47.36	30.92
500000	46.89	30.44
800000	46.20	29.73
1000000	46.02	29.52

5 lentelė. Maišytumas ir NŽ 65 tūkst. žodynui(anglų ir rusų k. pagal E.Whittaker[5])

Kalba	Maišytumas	NŽ %
Lietuvių	414.3	10.92
Rusų	387.4	7.62
Anglų	216.1	1.10

3.3.2. Sudėtiniai statistiniai lietuvių kalbos modeliai

Naudodami (11) kalbos modelio išraiką ir skaidydami žodžius pirmuoju būdu (pagal galūnių sąrašą) bei antruoju būdu (Lemuoklio pagalba), gavome tokius rezultatus.

6 lentelė. Rezultatai dirbant su 1422746 žodžiais

Skaidymo funkcija	Prasminis modelis, P_m		Morfologinis modelis, P_m		Mišrus modelis, P_{me}	Bendras modelis, P	
	Žodyno dydis	Maišytumas	Žodyno dydis	Maišytumas	Maišytumas	Maišytumas	NŽ %
Pagal galūnių sąrašą	1371372	687.83	3	1.49	1.05	687.83	1.16
	505437	386.31	489	31.97	3.68	1423.86	0.5
Lemuoklis	734695	358.22	1445	45.49	3.77	1350.7	0.84

Visias atvejais, geriausia gauta λ reikšmė – $\lambda = 0$. T.y. morfologinės žodžių informacijos 3-grama skaičiuojant bendrą modelio maišytumą nebuvo naudojama.

5. Diskusija

Šis darbas – tai pirminis mūsų bandymas sukurti ir įvertinti įvairius lietuvių kalbos statistinius modelius. Eksperimentai parodė, kad gaunami lietuvių kalbos statistiniai modeliai yra blogesni lyginant su anglų ir su rusų kalbų modeliais. Didesnį lietuvių kalbos modelių maišytumą galime bandyti paaiškinti tuo, kad mūsų naudojamas tekstynas yra labai nevienalytis, tuo tarpu kai anglų ir rusų kalbos modeliai kurti labiau vienalyčių tekstynų pagrindu. Priežastis gali būti ir sakinių pažymų nebuvimas mūsų naudotame tekstyne. Akivaizdu, kad žodžių priklausomybė sakinyje yra didesnė, negu tarp skirtingų sakinių ir, žinant informaciją apie sakinio ribas, modelis taptų tikslesnis. Papildoma priežastis – rašybos klaidos, atsiradusios arba teksto rinkimo metu, arba dėl simbolių kodavimo nevienodumo. Klaidų skaičius yra viena iš priežasčių, dėl ko mūsų modelio žodynas yra daug didesnis lyginant su kitų autorių eksperimentuose aprašytų kalbų žodynais. Išsprendus aukščiau išvardintas problemas, KM maišytumo ir NŽ įvertis turėtų pagerėti.

Iš dalies nepasitvirtino mūsų pasiūlytas (10) išraiškos supaprastinimas, keičiant ją (11) išraiška. Optimizuojant λ parametrą su optimizavimui skirta imtimi, gaudavome kraštutinį atvejį $\lambda = 0$, kuris faktiškai atmesdavo morfologinių žodžio dalių 3-gramos modelį. Morfologinės žodžio dalies prognozė, žinant prasminę žodžio dalį pasirodė esanti daug tikslesnė, nei žinant dviejų prieš ją esančių žodžių morfologinę informaciją. Kaip įjungti morfologinį modelį į bendrą kalbos modelį, lieka tolesnio tyrimo uždavinys. Galbūt atsakius į šį klausimą, galėtume sumažinti ir maišytumo įvertį, kuris mūsų nedžiugina, nes yra kelis kartus didesnis už paprastos 3-gramos įvertį.

[7] Vienetinė – kuri mokymo tekste pasitaikė 1 kartą.

Daug vilčių teikia morfologinis žodžių analizavimas. Mūsų naudotas Lemuoklis šiuo metu nesugeba apdoroti visų tikrinių daiktavardžių, kurių tekste yra tikrai daug. Jei būtų apdorojama bent didesnė dalis šių žodžių – tikėtina, kad žymiai sumažintume pagrindinių formų žodyno dydį, o tuo pačiu ir pagrindinių formų 3-gramos maišytumą.

Tolesnės šio darbo kryptys galėtų būti surasti būdą, kaip įjungti galūnių modelį į bendrą kalbos modelį, išnagrinėti kalbos modelį paremtą žodžių, pagrindinių formų ir kamienų grupavimu į klases.

6. Išvados

Šiame darbe įvertinti statistinio lietuvių KM maišaties ir NŽ parametrai. Su 65000 žodynu maišatis gauta 414.3, NŽ sudarė 10.92%. Dirbant net su 1.4 mln. žodynu NŽ buvo lygus 1.21%, t.y. nebuvo pasiekta 1% NŽ riba. Iš dalies tai galima paaiškinti tuo, kad mūsų turimas tekstynas yra nevienalytis. Jis skirtas atspindėti visai kalbai, o ne konkrečiai jos vartojimo sričiai. Iš kitos pusės, tai rodo, kad lietuvių kalboje ir žodynas, ir žodžių kaitomumo lygis yra didelis. Todėl remiantis šiais rezultatais, galima teigti, kad sprendžiant lietuvių šnekos atpažinimo uždavinį reikią ieškoti kitų kalbos dėsningumus aprašančių modelių, negu standartinis 3-gramos KM.

Mūsų pasiūlyta žodžių skaidymo ir tikimybių įvertinimo idėja nauja tuo, kad tai buvo bandymas atskirti prasminę ir morfologinę informaciją tekste, bei įvertinti žodžio tikimybę naudojant du atskirus abiejų tipų informacijos statistinius modelius. Kaip vieną iš šio būdo privalumų, galime pažymėti tai, kad pavyko sumažinti NŽ net iki 0.5%, ko neįmanoma padaryti lietuvių kalbai su standartinę 3-grama.

Palyginus skaidymo funkcijas - pagrįstą galūnių sąrašu ir pagrįstą morfologinę analize, išryškėja pastarojo būdo privalumas.Tai, kad pagrindinių formų 3-gramos modelio maišytumas geresnis už žodžių kamienų 3-gramos modelį, net esant didesniam pagrindinių formų žodynui, rodo morfologinės analizės perspektyvas.

Literatūros sąrašas

[1] **D. Carter, J. Kaja, L. Neumeyer, M. Rayner, F. Weng, M. Wiren**. Handling Compound Nouns in a Swedish Speech-Understanding System. *ICSLP 96*, Philadelphia, 1996.

[2] **P. Clarkson and R. Rosenfeld**. Statistical Language Modeling Using the CMU-Cambridge Toolkit, *EUROSPEECH 97,* Rhodes, Greece, 1997.

[3] **D. Jurafsky, J.H. Martin**. *Speech and Language Processing*, Prentice – Hall, New Jersey, 2000.

[4] **C. Martins, J.P. Neto, L.B. Almeida**. Using Partial Morphological Analysis in Language Modeling Estimation for Large Vocabulary Portuguese Speech Recognition. *Eurospeech 1999*, Budapest, Hungary, 1999.

[5] **E.W.D. Whittaker**. *Statistical Language Modelling for Automatic Speech Recognition of Russian and English*. PhD thesis, Cambridge University, Cambridge, 2000.

[6] **V. Zinkevičius**. Lemuoklis – morfologinei analizei. *Darbai ir dienos*, Kaunas, 2000.

Statistical modelling of Lithuanian language

This paper describes the work for creation of statistical models for Lithuanian language. We present 3 experiments. First we evaluate a performance for standard 3-gram model. Also we introduce two new types of statistical language models. One is based on the word parsing into the main part of the word and ending. A second model uses morphological analysis for word parsing. We evaluate and present the perplexity and OOV rate for these models depending on vocabulary size.