KOMPIUTERINĖ LINGVISTIKA/ COMPUTATIONAL LINGUISTICS

Dažniausios lietuvių kalbos morfologinio daugiareikšmiškumo rūšys ir jų automatinis vienareikšminimas

Erika Rimkutė, Aušra Grybinaitė

Anotacija. Straipsnyje pristatomas mažai tyrinėtas lietuvių kalbos morfologinis daugiareikšmiškumas ir pirmieji automatinio vienareikšminimo bandymai. Rašoma apie automatiniu būdu sulemuotą ir morfologiškai anotuotą lietuvių kalbos tekstyną. Ištyrus anotuotą tekstyną, paaiškėjo, kad kalba yra labai daugiareikšmė – apie 50 proc. žodžių ar žodžių formų yra morfologiškai daugiareikšmės. Norint panaudoti morfologiškai anotuoto tekstyno duomenis tolesniems tyrimams (automatinei sintaksinei analizei, kelių kalbų lygiagrečiam nagrinėjimui, automatiniam vertimui) reikia turėti vienareikšmes formas. Kitoms kalboms yra sukurta nemažai specialių vienareikšminimo programų. Lietuvoje ši sritis yra visai nauja ir mažai tirta, todėl tik visai neseniai buvo pradėta išsamiau gilintis į morfologinį daugiareikšmiškumą ir jo ribojimo galimybes. Šis straipsnis – tai lingvistų ir informatikų bendradarbiavimo, ribojant morfologinį daugiareikšmiškumą, rezultatų aptarimas.

Įvadas

Lietuvių kalba gerokai atsilieka nuo kai kurių kalbų kompiuterizavimo. Kitoms kalboms yra sukurtos ir nuolatos tobulinamos automatinės kalbos analizės programos: morfologijos, sintaksės, semantikos. Pavyzdžiui, Prahos priklausomybių medžių bankas yra pažymėtas morfologiškai, analitiškai sintaksiškai ir tektogramatiškai (Hajičová, 1998). Rusų priklausomybių medžių bankas yra sulemuotas, anotuotas morfologiškai ir sintaksiškai (Boguslavsky *et al.*, 2000). Anglų ir vokiečių kalbos gali pasigirti net keletu įvairiai anotuotų tekstynų (pavyzdžiui, plg. Brants *et al.*, 2001).

Lietuvoje jau irgi yra keletas tekstynų. Vienas iš jų – tai Vytauto Didžiojo universiteto Kompiuterinės lingvistikos centre daugiau nei iš 100 mln. žodžių sudarytas ir nuolatos pildomas „Dabartinės lietuvių kalbos tekstynas“ (adresas: http://donelaitis.vdu.lt/Tekstynas). Taip pat jau šis tas padaryta ir automatinės analizės labui: jau turime 1 mln. žodžių automatiniu būdu sulemuotą ir morfologiškai anotuotą tekstynėlį.

Automatinis teksto lemavimas ir morfologinis anotavimas – tai pirmosios automatinės kalbos analizės pakopos. Iš karto po tokios analizės išryškėja ne tik galimybė kalbą analizuoti greičiau, objektyviau, formaliau, bet ir trūkumai, iš kurių pagrindinis – kalbos daugiareikšmiškumas. Su šia problema susidūrė visų kalbų, turinčių tekstynus, tyrėjai, bandantys automatiškai analizuoti kalbas, nepaisant to, kad tos kalbos yra kitokios struktūros.

Pastaruoju metu daug dėmesio skiriama daugiareikšmiškumo problemų sprendimui (angl. *ambiguity resolution*). Kuriami įvairūs įrankiai, galintys riboti morfologinį, sintaksinį, semantinį daugiareikšmiškumą, sudarinėjamos taisyklės, padedančios vienareikšminti tekstus. Lietuviai šiuo atžvilgiu gerokai atsilieka, bet tai suprantama, nes visai neseniai buvo pradėti kurti anotuoti lietuvių kalbos tekstynai.

Vis dėlto pirmieji darbai jau padaryti: išanalizuotas ir suklasifikuotas lietuvių kalbos morfologinis daugiareikšmiškumas (toliau *morfologinis daugiareikšmiškumas* žymimas MD); jau žinoma, kokios formos, morfologinės kategorijos yra pačios daugiareikšmiškiausios, todėl galima pradėti kurti vienareikšminančias programas. Šiame straipsnyje pristatysime anotuoto tekstynėlio kūrimą, tvarkymą, lietuvių kalbos morfologinį daugiareikšmiškumą ir jo ribojimą pritaikant lingvistines, statistines bei kitokio pobūdžio taisykles.

Anotuotas lietuvių kalbos tekstynas

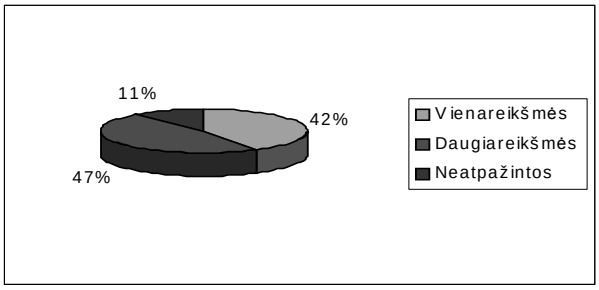
Vytauto Didžiojo universiteto Kompiuterinės lingvistikos centre 1 mln. žodžių automatiškai anotuoto tekstynėlio sudarymui buvo naudojama V. Zinkevičiaus sukurta kompiuterinė programa „Lemuoklis“ (Zinkevičius, 2000), galinti pateikti antraštinį rašytinės žodžio formos pavidalą, vadinamąją lemą, pvz., forma *laužo* sulemuojama kaip daiktavardis *laužas* arba kaip veiksmažodis *laužyti*. „Lemuoklis“ taip pat gali nurodyti ir morfologines pažymas. Minėta forma būtų apibūdinta kaip vyriškosios giminės daiktavardžio vienaskaitos kilmininkas (***laužo*** *liepsna*) arba tiesioginės nuosakos esamojo laiko veiksmažodžio trečiasis asmuo (*vaikai* ***laužo*** *šakas*).

Tekstyno žymėjimą galima suskirstyti į tris etapus: pirmiausia grynas tekstas lemuojamas – pateikiama tekstyne pavartoto žodžio antraštinė forma, t. y. lema. Antrojo etapo metu gali būti pateiktas žodžio formos morfologinis apibūdinimas. Toliau turėtų būti nustatomi kelis antraštinius pavidalus turintys žodžiai, t. y. vienareikšminama (plg. ankstesnį pavyzdį: ***laužo*** *liepsna – vaikai* ***laužo*** *šakas*). Tam tikslui reikalinga speciali programa, kurios pagrindinė funkcija būtų dviprasmybių panaikinimas (Zinkevičius, 2000; Marcinkevičienė, 2000).

Anotuotas tekstynėlis sudarytas iš įvairaus pobūdžio tekstų. Jame stengtasi išlaikyti didžiojo tekstyno proporcijas, todėl didžiąją dalį sudaro publicistikos tekstai. Pagal atitinkamas proporcijas anotuotame tekstynėlyje įdėta grožinės literatūros, mokslinio, administracinio stiliaus tekstų. Stengtasi apimti kuo įvairiausias kalbos atmainas, todėl į šį tekstynėlį pateko netgi Lietuvos Respublikos Seimo stenogramų, kurios artimiausios šnekamajai kalbai.

Anotuoti tekstynai dažnai būna nedideli – kelių šimtų tūkstančių ar kelių milijonų žodžių. Pavyzdžiui, vokiečių anotuotas *TIGER* tekstynas yra tik milijono žodžių (CS, 2001), čekų kalbos anotuoti tekstynai dažnai būna kelių šimtų tūkstančių žodžių (Hajič *et al.*, 1998; Hladká, 2000). Rusų kalbos priklausomybių medžių banką sudaro 1 milijonas žodžių (Boguslavsky *et al.*, 2000). Tokios pačios apimties yra ir minėtas lietuvių kalbos tekstynėlis.

Automatiškai sulemuoti ir morfologiškai pažymėti tekstai buvo peržiūrėti lingvisto, kuris panaikino daugiareikšmius atvejus, nes apie 47 proc. visų žodžių ar žodžių formų yra morfologiškai daugiareikšmės (žr. 1 paveikslą), įrašė neatpažintų ar klaidingai atpažintų žodžių lemas ir morfologines pažymas.



1 paveikslas. Vienareikšmių ir daugiareikšmių žodžių / žodžių formų santykis

Morfologinis daugiareikšmiškumas

Morfologiškai daugiareikšmiais vadinami tie žodžiai ar žodžių formos, kuriems „Lemuoklis“ automatiškai nustato dvi ar daugiau lemų arba kuriems pateikia dvi ar daugiau galimų morfologinių pažymų, pvz., *namo* – daiktavardis ir prieveiksmis (***nãmo*** *stogas – einu* ***namõ***), *galimas* – būdvardis ir dalyvis (***galimas*** *daiktas*), *laimės* – daiktavardis ir veiksmažodis (*lãimės kūdikis – jis* ***laimẽs*** *varžybas*) ir pan. (žr. 2 paveikslą).

Anotuotas netvarkytas tekstas	Eksperto sutvarkytas tekstas
Mane įvrd <aš> įvrd vnsk G **žavi** bdvr <žavus> bdvr teig nelygin.l neįvardž mot.gim vnsk V vksm <žavėti(-i,-ėjo)> vksm nesngr tiesiog.nuos esam.l vnsk Iiasm vksm nesngr tiesiog.nuos esam.l vnsk IIIasm vksm nesngr tiesiog.nuos esam.l dgsk IIIasm **minkšta** bdvr <minkštas> bdvr nelygin.l neįvardž mot.gim vnsk V bdvr nelygin.l neįvardž mot.gim vnsk Įn bdvr nelygin.l neįvardž bevrd.gim **katės** dktv <katė> dktv mot.gim vnsk K dktv mot.gim dgsk V dktv mot.gim dgsk Š vksm <katėti(-a,-ėjo)> vksm nesngr tiesiog.nuos būs.l vnsk IIIasm vksm nesngr tiesiog.nuos būs.l dgsk IIIasm **eisena** dktv <eisena> dktv mot.gim vnsk V dktv mot.gim vnsk Įn dktv mot.gim vnsk Š	**Mane** įvrd <aš> įvrd vnsk G **žavi** vksm <žavėti(-i,-ėjo)> vksm teig nesngr tiesiog.nuos esam.l vnsk IIIasm **minkšta** bdvr <minkštas> bdvr teig nelygin.l neįvardž mot.gim vnsk V **katės** dktv <katė> dktv mot.gim vnsk K **eisena** dktv <eisena> dktv mot.gim vnsk V

2 paveikslas. Netvarkyto ir eksperto peržiūrėto morfologiškai anotuoto teksto palyginimas

Morfologinį daugiareikšmiškumą galima įvardyti kaip reiškinį, kuris apima: 1) kaitomas ir nekaitomas; 2) skirtingų ir tų pačių kalbos dalių; 3) tam tikromis formomis, prozodiniais elementais besiskiriančias ir visiškai sutampančias žodžių formas bei žodžius.

Šiuo metu vienintelė Lietuvoje automatinę morfologinę analizę atliekanti programa „Lemuoklis“ padėjo pastebėti dažną, bet iš konteksto nenustatomą daugiareikšmiškumą. Tik šios programos anotuoti tekstai gali būti MD analizės objektas. Pagrindinis trūkumas yra „Lemuoklio“ nesugebėjimas analizuoti konteksto. Ši programa nenaudoja ir neturi informacijos apie semantiką. „Lemuoklis“ žodžių reikšmes sugeba nustatyti ne naudodamas kokius nors žodžių formų sąrašus su nurodytais tų formų morfologiniais apibūdinimais, o turėdamas lietuviškų žodžių šaknų sąrašą. Prie kiekvienos šaknies nurodomi naudojami skaitmeniniai kaitybos ir darybos modeliai. V. Zinkevičiaus sukurta programa taip pat nenaudoja ir neturi informacijos apie žodžių ar žodžių formų dažnines charakteristikas. Be to, MD analizė yra gana specifinė: analizuojamos rašytinės, be kirčio ženklų žodžių formos, todėl tiriamos homoformos gali skirtis ir prozodiniais elementais, pavyzdžiui, ***gėlės*** *žiedas – žydi* ***gėlės***.

Dėl šių priežasčių atsiranda tokių daugiareikšmiškumo atvejų, su kuriais realiai susidurti beveik neįmanoma ir kurie atrodo visiškai nerealūs, pavyzdžiui, daiktavardžiai *padarytis*, *kokis*. „Lemuoklis“ automatiškai iš visų daiktavardžių sugeneruoja mažybines jų formas. Tai padaryta dėl paprastos priežasties: „Dabartinės lietuvių kalbos žodyne“ pateikta labai nedaug mažybinių formų ir tik tokių, kurios nuo pamatinio žodžio nutolusios semantiškai, pvz., *stiklas – stikliukas*. Kuriant morfologinės analizės programą „Lemuoklis“ pagrindinės leksikos žinios buvo imtos iš „Dabartinės lietuvių kalbos žodyno“ ir „Tarptautinių žodžių žodyno“, bet juose

dažniausiai pateikti tik pamatiniai žodžiai, pvz., nurodytas tik daiktavardis *stalas*, o kalbos vartotojai patys turi suprasti, kad iš jo galima sudaryti deminutyvą *staliukas*. Taigi norint, kad kuo daugiau formų būtų atpažįstama automatiškai, teko kai kuriuos darybos būdus įtraukti kaip reguliarius, neatsižvelgiant į semantinius skirtumus. Dėl to vėliau paaiškėjo, kad buvo sugeneruota nerealiųjų homonimų, kurie sutapdavo su realiais lietuvių kalbos žodžiais, pvz., jau minėto nerealus mažybinio daiktavardžio *padarytis* (padaryto iš daiktavardžio *padaras*) vienaskaitos šauksmininkas (*mano mažasis* ***padaryti!***) sutampa su veiksmažodžio *padaryti* bendratimi (*privalai tai* ***padaryti***).

Homoformų atsiranda ir dėl kitų priežasčių: pvz., „Dabartinės lietuvių kalbos žodyne“ pateiktas retai vartojamas daiktavardis *kokis*, reiškiantis „kokybė“. Šio žodžio vienaskaitos kilmininkas (*siekiame geresnio* ***kokio***) sutampa su labai dažnu lietuvių kalbos įvardžio *koks* forma *kokio* (***kokio*** *megztinio ieškai?*). Skaitantiems rišlų tekstą aišku, kada koks žodis pavartotas, bet „Lemuoklis“ analizuoja pavienius žodžius, visiškai nenagrinėdamas tekste pavartotų gretimų žodžių, nenustatinėja jų ryšių, todėl ir atsiranda tokių daugiareikšmių formų, kurias paprastam kalbos vartotojui sunku net įsivaizduoti.

Visi šie dalykai didina morfologinių formų atpažinimo klaidingumą, trukdo automatinei analizei. Nepaisant kai kurių trūkumų, morfologinės pažymos, kad ir ne visada tikslios, padeda kitaip analizuoti kalbą: pamatoma tai, kas atrodo visiškai neįmanoma, nors teoriškai teisinga.

Dažniausios MD rūšys ir jų ribojimas

Morfologiškai daugiareikšmės formos buvo analizuotos išsamiau. Buvo nustatyta, didžiąją daugiareikšmių formų dalį (76,2 proc.) sudaro kaitomų kalbos dalių homoformos (pvz., sutampa du daiktavardžiai *būda* ir *būdas*: *sukalė šuniui* ***būdą*** *– išsiugdė gerą* ***būdą***). Gana dažni nekaitomų kalbos dalių sutapimai (pvz., žodelis *ir* gali būti jungtukas, prieveiksmis ir dalelytė). Tai 17,5 proc. visų morfologiškai daugiareikšmių formų. Pačios rečiausios (6,3 proc.) – kaitomų ir nekaitomų kalbos dalių homoformos (pvz., sutampa veiksmažodis ir prieveiksmis: ***vėliau*** *plaukus –* *ateisiu* ***vėliau***).

Šios homoformos buvo nagrinėtos dar smulkiau – pagal kalbos dalių sutapimą suskirstytos į 43 rūšis (išsamesni MD tyrinėjimai pateikti Rimkutė, 2003).

Pačios dažniausios homoformų rūšys yra veiksmažodžių vienaskaitos ir daugiskaitos trečiųjų asmenų (pvz., *jis* ***eina*** *– jie* ***eina***), nekaitomų kalbos dalių sutapimas (*tik* – tai jungtukas, dalelytė, ištiktukas ir prieveiksmis), vardažodžių linksnių sinkretizmas (*jauna* (tai gali būti vienaskaitos vardininkas, įnagininkas ir bevardė giminė) *moteris* (tai gali būti vienaskaitos vardininkas arba daugiskaitos galininkas), sutampančios bendraties bei dalyvio formos (pvz., *turi tai* ***padaryti*** *– blogai* ***padaryti*** *darbai*) ir būdvardžių bei prieveiksmių sutapimas (***jaunai*** *mokytojai nesiseka dirbti – atrodai labai* ***jaunai***).

Homoformų vienareikšminimas būtinas kuriant automatinio vertimo programas, dviejų kalbų lygiagrečiam nagrinėjimui, kuriant šnekos generavimo iš teksto sistemas, kuriant rašybos ir gramatikos tikrinimo programas, pirminiam teksto apdorojimui prieš sintaksinę analizę, informacijos išrinkimui ir paieškai.

Riboti lietuvių kalbos MD nėra lengva. Pirmiausia reikia apsibrėžti teorines nuostatas: koks daugiareikšmis žodis laikomas kokia kalbos dalimi, kaip skirtinos homonimiškos formos ir pan. Pavyzdžiui, labai problemiškas veiksmažodžių trečiojo asmens skyrimas: ar skirti atskirai vienaskaitos ir daugiskaitos trečiąjį asmenį, ar ne, kaip pažymėti beasmenius veiksmažodžius.

Taip pat nelengva panaikinti nekaitomų kalbos dalių homoformas, nes gramatikos ir žodynai gana nenuosekliai aprašo šias kalbos dalis; kartais nematyti didelių skirtumų tarp nurodytų kelių to paties žodelio reikšmių. Taigi ne visada net lingvistas analizuodamas rišlų tekstą gali nuspręsti, koks nekaitomas žodelis pavartotas. Dėl to sudėtinga kurti taisykles, ribojančias nekaitomų žodelių daugiareikšmiškumą.

Sutampantys vardažodžių linksniai yra dažna ir nelengvai išsprendžiama problema. Bandyta nustatyti ir pritaikyti keletą taisyklių ir šiai daugiareikšmiškumo rūšiai, bet dar nespėta išsamiai išanalizuoti linksnių sinkretizmo. Tai bus padaryta ateityje.

Mažinant lietuvių kalbos MD jau prasidėjo lingvistų ir informatikų bendradarbiavimas: lingvistas kuria teorines taisykles, kaip galima skirti sutampančias formas, o informatikas pritaiko jas, ieško kitokių formalių kriterijų, kaip skirtinos homoformos ir visą šią informaciją pritaiko kompiuterinei programai.

Ribojant kitų kalbų MD gauti tokie rezultatai: anglų kalboje pasiektas maždaug 96–97 proc. tikslumas, rumunų – 98,5 proc., čekų – 96 proc., slovakų kalboje pasitaiko 6,4 proc. neteisingų variantų, jei analizuojami tik žinomi žodžiai (Hajič, 2000; Hladká, 2000).

Anotuojant vokiečių kalbos *NEGRA* tekstyną pasiekiamas 99,2 proc. tikslumas, kai klasifikuojamos labiausiai tikėtinos morfologinės pažymos (jų yra apie 85 proc.), bet gaunamas tik 83 proc. tikslumas, jei pažymų tikėtinumas yra mažas (tokių pažymų būna apie 15 proc.) (Brants *et al*., 2000; Brants *et al*., 2001).

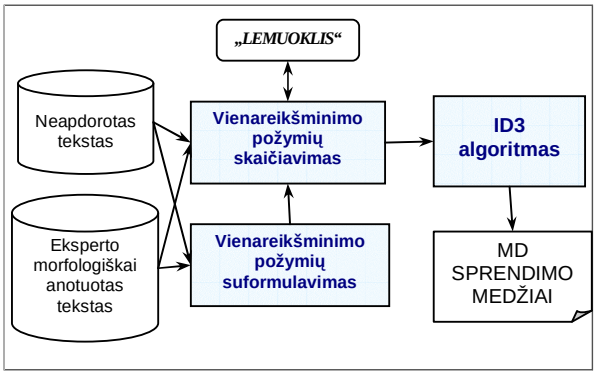
Vienareikšminimui pritaikomi įvairūs metodai, pavyzdžiui, čekų kalbos morfologiškai daugiareikšmės formos buvo analizuojamos ir vienareikšminamos remiantis statistiniais bei analitiniais metodais (Mírovský, 1998). Dažnai daugiareikšmiškumas ribojamas pritaikius slaptųjų Markovo modelių (HMM) analizę (Manning, 2000; Charniak, 1993). Šis metodas buvo pritaikytas Nacionaliniame britų tekstyne. Buvo pasiektas 96–97 proc. tikslumas, priklausomai nuo analizuojamo teksto rūšies (Leech *et al.*, 1994).

Ribojant lietuvių kalbos MD, buvo taikyti slaptieji Markovo modeliai, kurie yra tikimybiniai ir priklauso statistinio mokymosi modelių klasei. Daugiareikšmiškumo uždavinys buvo sprendžiamas iš dalies: bandyta nustatyti tik kalbos dalį, nenagrinėjant kitų morfologinių požymių. Naudotas Viterbi algoritmas, pagrįstas prielaida, kad kalbos dalis priklauso tik nuo prieš tai buvusios kalbos

dalies (bigramų atveju; trigramų atveju – nuo dviejų prieš tai buvusių kalbos dalių), t. y. nepriklauso nuo visos kalbos dalių sekos (Griciūtė, 2001; Griciūtė *et al.*, 2001).

Pastebėta, kad taikyti metodai turi kai kurių trūkumų. Pavyzdžiui, analizuojant morfologiškai daugiareikšmių formų kontekstą pastebėta, kad kalbos dalis gali priklausyti ne tik nuo prieš tai buvusios kalbos dalies (ar nuo dviejų kalbos dalių), bet homoformos ribojimą nulemiantis žodis gali būti nutolęs per daugiau kalbinių vienetų tiek prieš homoformą, tiek ir po jos. Sukurtoje vienareikšminimo sistemoje pritaikius Viterbi algoritmą randama tikėtiniausia sakinį atitinkanti kalbos dalių seka. Jei pasitaiko vienareikšminimo klaidų, sunku rasti ir pašalinti priežastį, dėl kurios atsiranda klaidų.

Pastaruoju metu MD vienareikšminimui taikomas ID3 algoritmas (žr. 3 paveikslą). Pasirinktas algoritmas sukuria medžius, kurie gana nesudėtingai gali būti perrašomi į taisyklių aibę, suprantamą tiek informatikui, tiek lingvistui.

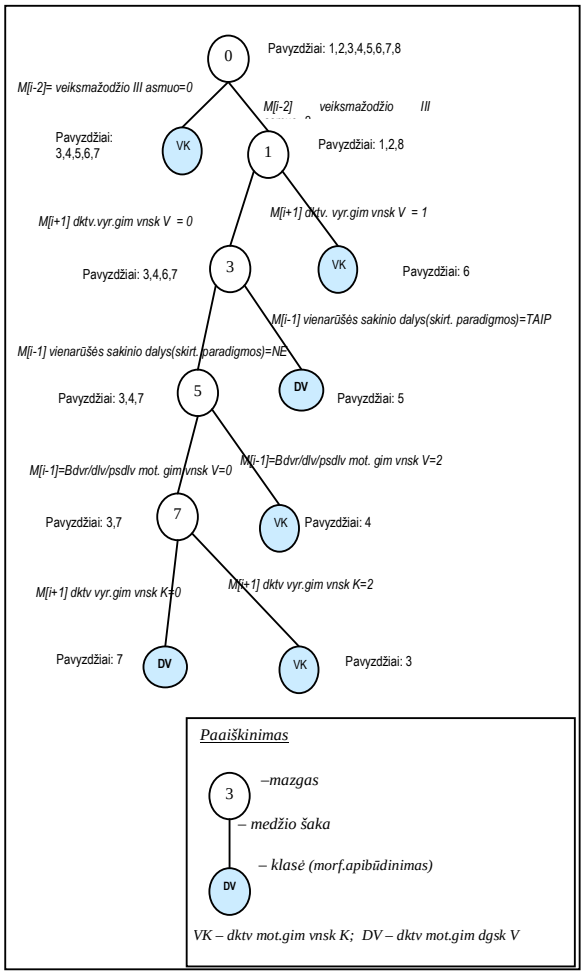


3 paveikslas. MD automatinio ribojimo schema

Sakiniuose ieškoma atskirų vienareikšminimo požymių kiekvienam daugiareikšmiškumo tipui. Vienareikšminimo požymiai neturi būti taisyklės – taisykles sukurs ID3 algoritmas, atrinkdamas geriausiai daugiareikšmes formas klasifikuojančius požymius ir iš jų sukurdamas neretai sudėtingos struktūros MD sprendimo medžius (plačiau žr. Grybinaitė, 2003).

Pagrindinis ID3 algoritmo veikimo principas: surasti vienareikšminimo požymį geriausiai klasifikuojantį mokymo imtį (nagrinėjamos MD rūšies pavyzdžius). Šis parametras naudojamas kaip medžio šaknis. Šis veiksmas kartojamas kuriant kiekvieną sprendimų medžio šaką. ID3 algoritmas sukuria sprendimų medį, kurio lapas žymi klasę (morfologinį apibūdinimą), o mazgas nusako, kad turi būti atlikta tolesnė analizė, kurios išdava būtų viena medžio šaka vienai parametro (diskretizuoto vienareikšminimo požymio) reikšmei.

Sprendimų medis koduoja vienareikšminimo taisykles. Kiekvienai MD rūšiai kuriamas atskiras sprendimų medis. 4 paveiksle pateikiamas sprendimų medis, sukurtas iš 8 MD pavyzdžių. Pateiktame medyje sukurtos taisyklės tik iš keleto pavyzdžių, todėl taikomos konkrečiam MD tipui. Sprendimų medžių, sukuriančių efektyvias vienareikšminimo taisykles, kūrimui reikalinga didelė homoformų pavyzdžių aibė.



4 paveikslas. Sprendimų medis, sukurtas iš 8 pavyzdžių

Išvados

Pritaikius lingvistines taisykles ir pastebėtus formalius MD ribojimo požymius, šiuo metu pasiekti tokie rezultatai:

1) išspręsta 73,65 proc. moteriškosios giminės vienaskaitos kilmininko ir daugiskaitos vardininko sinkretizmo atvejų;

2) išspręsta 81,65 proc. moteriškosios giminės vienaskaitos vardininko ir įnagininko sinkretizmo atvejų;

3) išspręsta 72,39 proc. veiksmažodžio trečiojo asmens (vnsk III asm, dgsk III asm ir III asm) sutapimo atvejų;

4) išspręsta 92,15 proc. bendraties ir neveikiamosios rūšies būtojo laiko neįvardžiuotinių vyriškosios giminės dalyvių daugiskaitos vardininko sutapimo atvejų;

5) išspręsta 42,52 proc. jungtukų, prieveiksmių ir dalelyčių sutapimo atvejų.

Visos šios grupės sudaro beveik 40 proc. viso MD. Kol kas teisingai pavyko vienareikšminti apie 25 proc. homoformų. Pateikti rezultatai gauti, kai morfologinio daugiareikšmiškumo sprendimų medžiams reikalaujamas

95 proc. lapo grynumas. Tačiau sukurti sprendimų medžiai neteisingai klasifikuoja 0,4 proc. homoformų. Jei kuriant sprendimo medžius pasirenkamas 100 proc. lapo grynumas, neteisingai suklasifikuotų homoformų nebelieka, tačiau ribojama tik apie 17 proc. lietuvių kalbos MD.

Ateityje planuojama toliau tirti daugiareikšmes formas, tobulinti jau esamas vienareikšminančias taisykles ir kurti naujas, kurios apims daugiau MD rūšių.

Literatūra

1. Boguslavsky, I., Grigorieva, S., Grigoriev, N., Kreidlin, L., Frid, N. (2000). Dependency Treebank for Russian: Concept, Tools, Types of Information. – In 18th International Conference on Computational Linguistics COLING-2000, Saarbrücken, Germany – http://acl.ldc.uppen.edu/C/C00/C00-2143.pdf.
2. Brants, T., Plaeh, O. (2000). Interactive Corpus Annotation. – *In Second International Conference on Language Resources and Evaluation LREC-2000*, May 31–June 2, 2000, Athens, Greece – http://www.coli.uni-sb.de/~plaehn/papers/lrec2000.pdf.
3. Brants, S., Dipper, S., Hansen, S., Lezius, W., Smith, G. (2001). The TIGER Treebank. – In Third Workshop on Linguistically Interpreted Corpora LINC-2001, Leuven, Belgium – http://www.coli.uni-sb.de/~sabine/tigertreebank.pdf.
4. Charniak, E. (1993). Statistical Language Learning. Cambridge, Massachusetts, London: A Bradford Book The MIT Press.
5. CS – Computerlinguistik und Sprachtechnologie. Eine Einführung (2001). Herausgeber K.-U. Karstensen, Ch. Elbert, C. Endriss *et al.* Heidelberg, Berlin: Spektrum Akademischer Verlag.
6. Grybinaitė, A. (2003). Dažniausių lietuvių kalbos morfologinių daugiaprasmybių sprendimas, taikant ID3 algoritmą. – VDU Informatikos fakulteto bakalauro darbas.
7. Griciūtė, V. (2001). Morfologinio daugiareikšmiškumo problemų sprendimas. Statistinio modelio realizacija taikant Viterbi algoritmą. – VDU Informatikos fakulteto bakalauro darbas.
8. Griciūtė, V., Pajarskaitė, G. (2001). Paslėptų Markovo modelių (HMM) taikymas morfologinio daugiareikšmiškumo problemos sprendimui. – Konferencijos „Informacinė visuomenė ir universitetinės studijos“ pranešimo medžiaga.
9. Hajič, J. (2000). Morphological Tagging: Data vs. Dictionaries. – In Proceedings of ANLP–NAACL Conference, pp. 94–101, Seattle, Washington, USA. – http://shadow.ms.mff.cuni.cz/pdt/Morphology_and_Tagging/Tagging/Doc/References/naacl00.pdf.
10. Hajič, J., Hladká, B. (1998). Tagging Inflective Languages: Prediction of Morphological Categories for a Rich, Structured Tagset. – ACL–Coling'98, Montreal, Canada, pp. 483–490. – http://acl/ldc.uppen.edu/P/P98/P98-1080.pdf.
11. Hajičová, E. (1998). Prague Dependency Treebank: From analytic to tectogrammatical annotations. – In Proceedings of the First Workshop on Text, Speech, Dialogue, pp. 45–50, Brno, Czech Republic. – http://shadow.ms.mff.cuni.cz/pdt/Corpora/PDT_1.0/References/Brno98.pdf.
12. Hladká, B. (2000). *Czech Language Tagging* – Doctoral thesis. Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Prague. http://shadow.ms.mff.cuni.cz/pdt/Morphology_and_Tagging/Tagging/Doc/References/disertace.pdf.
13. Leech, G., Garside, R., and Bryant, M. (1994). The large-scale grammatical tagging of text: Experience with the British National Corpus. – Corpus-based research into Language. Amsterdam-Alanta, GA: Radopi.
14. Manning, Ch. (2000). Probabilistic Models in Computational Linguistics – http://nlp. stanford.edu/~manning/talks/ima2000.pdf.
15. Marcinkevičienė, R. (2000). Tekstynų lingvistika (teorija ir praktika). – Darbai ir Dienos. T. 24, 7–64.
16. Mírovský, J. (1998). Morfologické značkování textu: automatická disambiguace. – Diplomová práce. Matematicko-fyzikálný fakulta, Univerzity Karlovy, Praha. http://shadow.ms.mff.cuni.cz/pdt/Morphology_and_Tagging/Tagging/Doc/References/jm.pdf.
17. Rimkutė, E. (2003). Morfologinio daugiareikšmiškumo tipologija. – Lituanistica, Nr. 4 (56), P. 60–78.
18. Zinkevičius, V. (2000). *Lemuoklis* – morfologinei analizei. – Darbai ir dienos. T. 24, 246–273.

Erika Rimkutė, Aušra Grybinaitė

The Most Frequent Types of the Morphological Ambiguity of the Lithuanian Language and the Automatical Disambiguation of Them

Summary

The article researches the morphological ambiguity, which was analysed in automatically tagged corpus of the Lithuanian language. The corpus with morphological tags has shown a large ambiguity of the language: almost 50 percent of word forms are ambiguous. The most frequent types of ambiguities are syncretism of singular and plural of the third person verbs, syncretism of non inflected parts-of-speech and case syncretism of nouns. This article presents linguistic, statistical rules and algorithms, that were created for morphological disambiguation. The constraints of disambiguation have been implemented in a programme that calculates the attributes and creates the learning set, necessary for creating the decision trees. For the meantime we have analysed about 40 percent of homoforms and have achieved 25 percent accurateness in disambiguation.

Straipsnis įteiktas 2004 01
Parengtas spaudai 2004 06

Apie autores

Erika Rimkutė, Vytauto Didžiojo universiteto Lietuvių kalbos katedros doktorantė.
Interesų sritys: tekstynų ir kompiuterinė lingvistika, automatinė morfologinė analizė, morfologinis daugiareikšmiškumas ir jo vienareikšminimas.
Publikacijos:
Rimkutė E. Morfologinio daugiareikšmiškumo tipologija. – Lituanistica, 2003, Nr. 4 (56), P. 60–77.
Rimkutė E. Homoformos dabartinės lietuvių kalbos tekstyne. – Lituanistica, 2002, Nr. 2 (50), P. 86–101.
Rimkutė E. *Lietuvių kalbos tekstyno morfologinės analizės automatizavimas.* – Informacinė visuomenė ir universitetinės studijos. 6-toji magistrantų ir doktorantų konferencija. Vytauto Didžiojo universitetas, Kauno technologijos universitetas, Vilniaus universitetas, 2001, P. 60–66.
Projektai: Valstybinės lietuvių kalbos komisijos projektas „Lietuvių kalbos automatinis morfologinis atpažinimas ir sintezė“; Valstybinės lietuvių kalbos komisijos projektas „Lietuvių kalba informacinėje visuomenėje 2000–2006“ tęstinio lietuvių kalbos tekstyno rengimui.
Adresas: Vytauto Didžiojo universiteto Kompiuterinės lingvistikos centras, K. Donelaičio g. 52–206, Kaunas.
El. paštas: Erika_Rimkute@fc.vdu.lt

Aušra Grybinaitė, Vytauto Didžiojo universiteto Informatikos fakulteto verslo informatikos specialybės magistrantė.
Interesų sritys: kompiuterinė lingvistika, morfologinis daugiareikšmiškumas ir jo automatinis vienareikšminimas.
Adresas: Vytauto Didžiojo universitetas, Vileikos g. 8, Kaunas.
El. paštas: Ausra_Grybinaite@fc.vdu.lt