VYTAUTO DIDŽIOJO UNIVERSITETAS

Vidas DAUDARAVIČIUS

TEKSTO SKAIDYMAS PASTOVIŲJŲ JUNGINIŲ SEGMENTAIS

Daktaro disertacijos santrauka
Fiziniai mokslai, informatika (09 P)

Kaunas, 2012

Disertacija rengta 2008–2012 metais Vytauto Didžiojo universitete

Mokslinis vadovas:
doc. dr. Minija Tamošiūnaitė
Vytauto Didžiojo universitetas, fiziniai mokslai, informatika – 09 P

Disertacija ginama Vytauto Didžiojo universiteto Informatikos mokslo krypties taryboje:

Pirmininkas:
prof. habil. dr. Vytautas Kaminskas
Vytauto Didžiojo universitetas, fiziniai mokslai, informatika – 09 P

Nariai:
prof. habil. dr. Laimutis Telksnys
Vytauto Didžiojo universitetas, fiziniai mokslai, informatika – 09 P
prof. dr. Rimantas Butleris
Kauno technologijos universitetas, fiziniai mokslai, informatika – 09 P
prof. dr. Ineta Dabašinskienė
Vytauto Didžiojo universitetas, humanitariniai mokslai, filologija – 04 H
dr. Rytis Maskeliūnas
Kauno technologijos universitetas, fiziniai mokslai, informatika – 09 P

Oponentai:
doc. dr. Vytautas Rudžionis
Vilniaus universitetas, fiziniai mokslai, informatika – 09 P
doc. dr. Violeta Kalėdaitė
Vytauto Didžiojo universitetas, humanitariniai mokslai, filologija – 04 H

Disertacija bus ginama viešame Informatikos mokslo krypties tarybos posėdyje 2013 m. sausio 31 d. 11 val. Informatikos ir gamtos mokslų bibliotekos Vinco Čepinskio skaityklo je
Adresas: Vileikos g. 8–605, Kaunas, Lietuva

Disertacijos santrauka išsiuntinėta 2012 m. gruodžio 31 d.
Disertaciją galima peržiūrėti Vytauto Didžiojo universiteto, Vilniaus universiteto Matematikos ir informatikos instituto ir Nacionalinėje Martyno Mažvydo bibliotekose

VYTAUTAS MAGNUS UNIVERSITY

Vidas DAUDARAVIČIUS

COLLOCATION SEGMENTATION FOR TEXT CHUNKING

Summary of Doctoral Dissertation
Physical Sciences, Informatics (09 P)

Kaunas, 2012

This doctoral dissertation was written at Vytautas Magnus University in 2008–2012

Research supervisor:

doc. dr. Minija Tamošiūnaitė
Vytautas Magnus University, Physical sciences, Informatics – 09 P

The dissertation will be defended at the Council of Informatics of Vytautas Magnus University:

Chair:

prof. habil. dr. Vytautas Kaminskas
Vytautas Magnus University, Physical sciences, Informatics – 09 P

Members:

prof. habil. dr. Laimutis Telksnys
Vytautas Magnus University, Physical sciences, Informatics – 09 P
prof. dr. Rimantas Butleris
Kaunas University of Technology, Physical sciences, Informatics – 09 P
prof. dr. Ineta Dabašinskienė
Vytautas Magnus University, Humanities, Philology – 04 H
dr. Rytis Maskeliūnas
Kaunas University of Technology, Physical sciences, Informatics – 09 P

Opponents:

doc. dr. Vytautas Rudžionis
Vilnius University, Physical sciences, Informatics – 09 P
doc. dr. Violeta Kalėdaitė
Vytautas Magnus University, Humanities, Philology – 04 H

The public defence of the dissertation is to be held in the meeting of the Council of Informatics of Vytautas Magnus University at 11 a.m. on January 31, 2013 in the Vincas Čepinskis athenaeum of Informatics and Natural sciences.

Address: Vileikos g. 8–605, Kaunas, Lithuania.

The summary of the dissertation was sent out to relevant institutions on 31 December 2012.

The dissertation is available at the library of Vytautas Magnus University Library, Library of Institute of Mathematics and Informatics, and Martynas Mažvydas National Library of Lithuania.

Santrauka

Teksto skaidymo įvairaus tipo segmentais metodai yra plačiai naudojami teksto apdorojimui. Tarp dažniausiai naudojamų teksto segmentavimo metodų yra teksto segmentavimas temomis, sakiniais, morfemomis, fonemomis ir kinų teksto segmentavimas žodžiais. Segmentuojant naudojami tiek statistiniai, tiek formalieji metodai. Disertacijoje pristatomas naujas segmentavimo tipas ir metodas – *segmentavimas pastoviaisiais junginiais*, ir pateikiami taikymai įvairiose teksto apdorojimo srityse:

– **Leksikografijoje.** Taikant pastoviųjų junginių segmentavimą atskleidžiama, kaip objektyviai ir greitai galima analizuoti labai didelius tekstų archyvus aptinkant vartojamą terminiją ir šių automatiškai identifikuotų terminų svarbumą ir kaitą laiko tėkmėje. Ši analizė leidžia greitai nustatyti svarbius metodologinius pokyčius mokslinių tyrimų istorijoje ir nustatyti pastarojo meto aktualias tyrimų sritis.

– **Tekstų klasifikavime.** Atskleidžiama, kaip taikant segmentavimą pastoviaisiais junginiais galima pagerinti tekstų klasifikavimo rezultatus. Tyrimu parodoma, kad segmentavimas pastoviaisiais junginiais ir nenaudojant jokių kitų automatinių kalbos analizės priemonių leidžia pasiekti geresnius rezultatus nei gaunami naudojant įvairias nuo kalbos priklausomas analizės priemones. ES teisės dokumentų klasifikavimo atveju rezultatai pagerinami nuo 48,30±0,48% iki 54,67±0,78%, nuo 51,41±0,56% iki 59,15±0,48% ir nuo 54,05±0,39% iki 58,87±0,36% atitinkamai anglų, lietuvių ir suomių kalboms.

– **Statistiniame mašininiame vertime.** Pasitelkiant segmentavimą pastoviaisiais junginiais atskleidžiama, kad nežymiai galima pagerinti statistinio mašininio vertimo kokybę, ir atskleidžiama įvairių žodžių junglumo įverčių įtaka segmentavimui pastoviaisiais junginiais.

  Naujas teksto skaidymo pastoviaisiais junginiais metodas atskleidžia naujas galimybes gerinti teksto apdorojimo rezultatus įvairiuose taikymuose ir įvairiose kalbose.

Summary

Segmentation is a widely used paradigm in text processing. There are many methods of segmentation for text processing, including: topic segmentation, sentence segmentation, morpheme segmentation, phoneme segmentation, and Chinese text segmentation. Rule-based, statistical and hybrid methods are employed to perform the segmentation. This dissertation introduces a new type of segmentation–*collocation segmentation*–and a new method to perform it, and applies them to three different text processing tasks:

– **Lexicography.** Collocation segmentation makes possible the use of large corpora to evaluate the usage and importance of terminology over time. It highlights important methodological changes in the history of research and allows actual research trends throughout history to be rediscovered. Such an analysis is applied to the ACL Anthology Reference Corpus of scientific papers published during the last 50 years in this research area.

– **Text categorization.** Text categorization results can be improved using collocation segmentation. The study shows that collocation segmentation, without any other language resources, achieves better results than the widely used n-gram techniques together with POS (Part-of-Speech) processing tools. The categorization precision of EU legislation using EuroVoc was improved, from 48.30±0.48% to 54.67±0.78%, from 51.41±0.56% to 59.15±0.48% and from 54.05±0.39% to 58.87±0.36% for English, Lithuanian and Finnish languages respectively.

– **Statistical Machine Translation.** The preprocessing of data with collocation segmentation and subsequent integration of these segments into a Statistical Machine Translation system improves the translation results. Diverse word combinability measures variously influence the final collocation segmentation and, thus, the translation results.

The new collocation segmentation method is simple, efficient and applicable to language processing for diverse applications.

Dažnai naudojami terminai ir akronimai

NLP – Natūralios kalbos apdorojimas (Natural Language Processing – angl.)

MI – Mutual Information

MT – Mašininis vertimas (Machine Translation – angl.)

TF–IDF – Term Frequency Inverse Document Frequency

SMT – Statisitnis mašininis vertimas (Statistical Machine Translation – angl.)

BLEU – Bilingual Evaluation Understudy

ACL – Kompiuterinės lingvistikos asociacija (Association for Computational Linguistics – angl.)

HMM – Paslėptasis Markovo modelis (Hidden Markov Model – angl.)

Turinys

Įvadas

"(Vienas) žodis nėra privilegijuotas prasmės perteikimo prasme. [...] Leksiniai vienetai gali būti pavieniai žodžiai, žodžių dūriniai, daugiažodžiai vienetai, frazės, ir taip pat idiomos" [Teubert, 2005]. Pastovusis junginys yra gerai žinomas kalbinis reiškinys, kurio tyrimų ir taikymų istorija yra ilga [Daudaravičius, 2012a]. Visgi, terminas *pastovusis junginys* neturi nusistovėjusio apibrėžimo. Šioje disertacijoje nesiekiama analizuoti teorinių pagrindų apie tai, *kas yra pastovusis junginys*. Yra esminis skirtumas tarp tų tyrėjų, kurie mano, kad pastovusis junginys yra svarbi teorinė sąvoka –

– "Pastoviojo junginio prasmė nėra tiesioginė prasminių jo dalių kompozicija. Pavyzdžiui, *red tape* prasmė yra visiškai kita nei jo sudedamųjų komponentų prasmės."[1] [Wermter and Hahn, 2004]

– "Pastovusis junginys yra dviejų ar daugiau žodžių junginys, kurie atitinka tam tikrą būdą įvardinti daiktus. Kartais, pastoviojo junginio sąvoka yra apibrėžiama sintaksės (pagal galimą kalbos dalių sekos modelį) arba semantikos (reikalaujanti pastovųjį junginį pateikti neskaidomą reikšmę) priemonėmis." [Franz and Milch, 2002]

– ir tų tyrėjų, kurie laiko pastoviuosius junginius patogiu deskriptyviu vienetu be jokio teorinio statuso:

– "Pastovusis junginys yra dviejų ar daugiau netoli vienas nuo kito tekste esančių žodžių reiškinys." [Frantzi and Ananiadou, 1996]

– "Pačia bendriausia prasme, pastovusis junginys yra įprastas arba leksikalizuotas žodžių derinys." [Weeds et al., 2004]

– "Pastovusis junginys yra pasikartojanti žodžių seka, kuri pasirodo dažniau nei atsitiktinai tam tikroje srityje." [Gil and Dias, 2003]

Lingvistai tiria pastoviuosius junginius leksiniu požiūriu. Nuo Melčuk [Melc'uk and Polguere, 1987] iki šių laikų kompiuterinės lingvistikos tyrėjų [Pecina, 2005] skiria didelį dėmesį pastoviojo junginio reiškinio apibrėžimui statistiniu ir sintaksiniu požiūriu.

Natūralios kalbos apdorojimo tyrėjai naudoja pastoviuosius junginius kaip esminę priemonę tokiuose uždaviniuose kaip natūralios kalbos struktūrinė analizė (pvz., sintaksinės ir semantinės struktūros analizė) ir sintezė (pvz., vertimas), o taip pat ir tokiuose praktiniuose taikymuose kaip mašininis vertimas, informacijos suradimas (retrieval -

[1] iš anglų kalbos vertė - Vidas Daudaravičius

angl.) ir ištraukimas (extraction - angl.). Dažnai *pastoviojo junginio* vartojimas yra susijęs su šiomis sąvokomis:

– **žodis** – bazinis kalbos vienetas. Žodžiai nėra paprasti reiškiniai. Apskritai, žodis yra dalinė ar pilna tam tikro reiškinio pasaulyje išraiška. Nors indo-europiečių kalbose žodžiai rašomi kaip atskirų raidžių sekos, Azijos šalių kalbose nėra aiškaus žodžio apibrėžimo.

– **leksema** – nedalomas elementas, paprastai tai žodis kalbos apdorojimo srityje. Žodžiavimas yra teksto skaidymas leksemų sekomis. Paprasti žodžiai yra lengvai atpažįstami, bet yra žodžių neturinčių aiškių ribų. Pavyzdžiui, leksema *don't* gali būti skaidoma kaip: vienas žodis | *don't* |, du žodžiai | *don* | *'t* | ar | *do* | *n't* |, arba trys žodžiai | *don* | *'* | *t* |. Nėra aiškių nuorodų kaip žodžiuoti tekstą leksemomis. Paprastai tai paliekama spręsti patiems inžinieriams. Nesenai atliktame tyrime [He and Kayaalp, 2006] parodyta, kad daug laisvai internete prieinamų žodžiavimo priemonių pateikia skirtingus rezultatus (vienuolika iš trylikos buvo skirtingi).

– n**-grama** – šalia einančių n žodžių seka tam tikrame tekste. Žodžių n-gramos yra plačiai paplitusi natūralios kalbos modeliavimo naudojant HMM (paslėptuosius Markovo modelius) priemonė. Jos taip pat naudojamos sudaryti daugiavektorines erdves, kurios reikalingos teksto klasifikavimui, mašininiam vertimui, informacijos paieškai, ir t. t.

Pastoviųjų junginių reiškinį yra ganėtinai paprasta suprasti, visgi sudėtinga pritaikyti realiuose uždaviniuose. Pastaraisiais dviem dešimtmečiais mažai pasistūmėta į priekį, kad pastoviuosius junginius būtų galima nesudėtingai pritaikyti įvairiuose uždaviniuose. [Smadja, 1991] pristatė bendrą trijų žingsnių metodą pastoviųjų junginių apdorojimui, ir parodė pastoviųjų junginių ištraukimo sitemos 80% tikslumą ir 94% aprėptį. Sistema buvo naudojama leksikografiniam uždaviniui – pastoviųjų junginių žodyno kompiliavimui, kurį sudarė apie 2,000 pastoviųjų junginių. Apskritai, pasiūlytas algoritmas leido surasti dviejų ir ilgesnių sekų žodžių pastoviuosius junginius. Nors rezultatai buvo sėkmingi ir rodė proveržį pastoviųjų junginių suradimo srityje, visgi šis pastoviųjų junginių metodas niekada nebuvo panaudotas kituose uždaviniuose. Pagrindinė kliūtis šio metodo taikymui buvo algoritmo sudėtingumas ir reikalingi skaičiavimo resursai. Net ir šiomis dienomis, taikymai, kuriuose galima būtų panaudoti tokią sistemą, yra mažai tikėtini dėl

per didelių skaičiavimo resursų. Todėl, be jokios rimtos priežasties, pastovieji junginiai buvo "užmiršti", ir ilgą laiką nebuvo pateikta naujų metodų. Tuo tarpu, n-gramos tapo pagrindine priemone kalbos apdorojimo uždaviniuose, nes nesudėtinga naudoti, paprasta pritaikyti, ir teoriškai pagrįstas. Pastovieji junginiai buvo toliau tiriami tik entuziastų. Praeito dešimtojo dešimtmečio pradžioje kompiuteriai tapo pakankamai galingi, kad galima būtų naudoti statistinius metodus (n-gramas), o lingvistinius metodus naudojantys formalieji (rule-based – angl.) metodai tapo per daug sudėtingi. Todėl, n-gramos buvo pasirinktos kaip pastoviųjų junginių pakaitalas natūralios kalbos apdorojimo srityje.

Pastoviųjų junginių paradigmos keitimo svarba yra keliama vienoje iš pastarųjų pastoviųjų junginių studijoje, [Seretan, 2011]. N-gramos neperteikia pagrindinės kalbos savybės – *prasmės reiškimo*, deskriptyvumo savybės. Kokio ilgio turi būti n-grama, kad išreikštų prasmę? Dažniausiai remiamasi principu, kad n-gramos turi būti tokio ilgio, kiek leidžia techniniai resursai ar apdorojimui skirto laiko ribos. Šis požiūris yra praktiškas, bet netinkamas lingvistiniu požiūriu. Praktikoje unigramos yra naudojamos, kai reikalingas greitis, o trigramos – kai reikalinga kokybė. Pagrindiniai sunkumai taikant pastoviuosius junginius yra nežinomas optimalus pastoviojo junginio ilgis ir kiek jis gali būti ilginamas.

Aukštesnės eilės n-gramoms reikalinga turėti didelius tekstynus patikimiems statistiniams duomenims išgauti. Tokie dideli tekstynai yra sudaryti plačiai paplitusioms kalboms, tokioms kaip: anglų, kinų, prancūzų, ir vokiečių. Pagrindinė problema tampa mažiau kalbinių išteklių turinčios kalbos. Nors didžioji dauguma tyrimu atliekama anglų kalbai, yra dar daug kitų kalbų, kurias reikalinga apdoroti. Naivu manyti, kad metodai, kurie sėkmingai taikomi anglų kalbai, tinka ir kitoms kalboms. Nėra daug daugiakalbių išteklių palyginamam kalbų apdorojimui. Didžiausias daugiakalbis visiems prieinamas tekstynas yra *The JRC–Acquis Multilingual Parallel Corpus*, sudarytas Europos Komisijos junginių tyrimų centre (Joint Research Center of the European Commission – angl.). Nors atsiranda vis daugiau tekstynų, tačiau trūksta kalbos apdorojimo priemonių, tokių kaip analizatoriai (parsers – angl.), mažiau kalbinių išteklių turinčioms kaboms.

Paskutinis svarbus klausimas norint pereiti nuo n-gramų prie pastoviųjų junginių yra: ar galima palyginti skirtingo ilgio pastoviuosius junginius? Pastaruoju metu bigramos, trigramos, 4-gramos ir 5-gramos yra surandamos atskirai. Be kita ko, nors pagrindinis tikslas yra prasmę turintys daugiažodžiai pastovieji junginiai, nėra palikta vietos vienažodžiams vienetams, t.y., vienažodžių (unigramų) vienetų statusas nėra aiškus, ir unigramos dažniausiai iškart yra atmetamos be jokio pagrindimo.

Šioje disertacijoje aš naudoju naują pastoviojo junginio apibrėžimą: **Pastovusis junginys yra ilgiausias nepersindengiantis žodžių junginys tekste, kuriame šalia einančių žodžių poros reiškiasi dažniau nei atsitiktinai**. Šis apibrėžimas praplečia tradicinį dažniausią pastoviojo junginio apibrėžimą (tokį kaip *Pastovusis junginys yra pasikartojanti žodžių seka, kuri pasirodo dažniau nei atsitiktinai.*) reikalavimu, kad pastoviųjų junginių apdorojimas vyksta tekste, o ne naudojant n-gramų filtravimą. Iš esmės, tai atitinka sakinio segmentavimo uždavinį: nors nėra paprasta apibrėžti *kas yra sakinys*, bet yra paprasta atpažinti sakinio pradžią ir pabaigą pradžioje net neatlikus jokios kitos analizės. Savo disertacijoje siekiu panaikinti aukščiau pateiktą tuštumą pateikiant naują segmentavimo tipą.

Disertacijos objektas yra žodžių pastovieji junginiai, kurie apjungia žodžių sekas į prasminius (pvz., *mašininis vertimas*) arba funkcinius (e.g., *be kita ko*) leksinius vienetus.

Disertacijos pagrindinis tikslas yra pateikti *segmentavimą pastoviaisiais junginiais*, pastoviųjų junginių apdorojimo metodą, kuris yra: mažo algoritminio sudėtingumo; neribojamas pastoviojo junginio ilgiu, nuo kalbos nepriklausomas (vien paprastas neanotuotas tekstas); efektyvus; ir pritaikomas įvairiuose natūralios kalbos apdorojimo uždaviniuose.

Pagrindiniai uždaviniai, kuriuos reikalinga išspręsti norint pasiekti užsibrėžtą tikslą yra:

– apibendrinti segmentavimo tipus, ir tokiu būdu pateikti pagrindinius uždavinius apibrėžiant segmentavimo vienetą (tekstas, diskursas, tema, pastraipa, sakinys, žodis, ir t.t.);

– aprašyti pastoviųjų junginių suradimo algoritmą, o taip pat sukurti bendrus segmentavimo pastoviaisiais junginiais principus, kurie apima įverčių, požymių ir slenksčių atranką, reikalingą pastoviųjų junginių suradimo sistemos įgyvendinimui;

– surasti optimalų junglumo, kuris nusako ryšio tarp dviejų žodžių stiprumą ir apibrėžia būtinybę apjungti šalia einančius žodžius vienu pastoviuoju junginiu, slenkstį;

– ištirti segmentavimo pastoviaisiais junginiais įtaką teksto klasifikavimo rezultatams skirtingoms kalboms (anglų, lietuvių, suomių), ir palyginti teksto klasifikavimo rezultatus naudojant unigramas ir bigramas;

– ištirti segmentavimo pastoviaisiais junginiais įtaką, palyginant su unigramomis, statistinio mašininio vertimui, taikant segmentavimą lygiagretiesiems tekstams;

– pritaikyti segmentavimą pastoviaisiais junginais ir ištirti pagrindines tyrimų kryptis analizuojant publikuotų straipsnių terminiją.

Disertacijos naujumas ir praktinė nauda:

– pateiktas naujas, nuo kalbos nepriklausomas, ir efektyvus pastoviųjų junginių apdorojimo metodas, kuris yra tinkamas daugelyje natūralios kalbos apdorojimo uždavinių, leidžia atiktrūkti nuo šiuo metu labiausiai taikomo n-graminio principo, ir yra praktikoje priimtino sudėtingumo.

– Ištirtas naujas pastoviųjų junginių apdorojimo metodas, pateikiant tekstą kaip junglumo tarp dviejų šalia einančių žodžių kreivę. Junglumo kreivė, kuri paremta dviejų šalia einančių žodžių junglumo įvertinimu, įgalina naujo metodo sukūrimą. Šis metodas leidžia išvengti būtinybės nustatyti slenkstį rankiniu būdu, kuris yra būtinas tradiciniame pastoviųjų junginių apdorojime.

– Išbandytos ir ištirtos naujos instrukcijos kaip taikyti pastoviuosius junginius tekstų klasifikavimui. Panašūs klasifikavimo rezultatai skirtingoms kalboms (taikant anglų, lietuvių ir suomių kalboms) rodo, kad segmentavimas pastoviaisiais junginiais leidžia pagerinti klasifikavimo rezultatus ir taikyti įvairioms kalboms. Segmentavimas pastoviaisiais junginiais leidžia pasiekti geriausius rezultatus lyginant su unigramomis ir bigramomis.

– Išbandytos ir ištirtos naujos instrukcijos kaip taikyti pastoviuosius junginius statistiniam mašininiam vertimui. Rezultatų analizė parodė, kad segmentavimas pastoviaisiais junginiais yra labai naudingas statistiniame mašininiame vertime tekstų lygiagretinimo etape.

– Išbandytos ir ištirtos naujos instrukcijos kaip taikyti pastoviuosius junginius terminijos vartosenos analizei dideliuose tekstynuose. Tyrimas parodė, kad segmentavimas pastoviaisiais junginiais leidžia nustatyti reikšmingus terminus, ir analizuoti šių terminų kasmetinį reikšmingumą.

Disertacijos struktūra. Skyriuje 2 yra pateikiama įvairių kalbos vienetų ir jų identifikavimo metodų apžvalga. Analizė rodo, kad pagrindinis platesnių vienetų identifikavimas remiasi ribų nustatymu. Aš panaudoju šią praktiką segmentavimo pastoviaisiais

junginiais metodui, kuris yra pateikiamas Skyriuje 3. Vėliau, Skyriuje 4, taikau segmentavimą pastoviaisiais junginiais ir analizuoju tris skirtingus uždavinius norėdamas nustatyti naujo metodo naudą. Svarbu pažymėti, kad segmentavimo pastoviaisiais junginiais integravimas statistiniame mašininiame vertime buvo atliktas bendradarbiaujant su Barcelona Media Innovation Center. Galiausiai, disertacijos pabaigoje yra pateikiamos pagrindinės išvados.

2 Segmentavimas pastoviaisiais junginiais

Segmentavimas pastoviaisiais junginiais yra naujas segmentavimo tipas, kurio tikslas yra surasti *pastovias žodžių sekas* ir segmentuoti tekstą žodžių sekomis, vadinamomis pastoviųjų junginių segmentais. Aš naudoju apibrėžimą *seka* kaip vienas ar daugiau. Todėl, pastoviojo junginio segmentas yra vieno ar daugiau žodžių seka su aukštu tarpusavio junglumu. Pastoviojo junginio segmentas gali būti bet kokio ilgio (net ir vienas žodis) ir ilgis nėra apibrėžiamas iš anksto. Šis apibrėžimas skiriasi nuo kitų pastoviųjų junginių apibrėžimų, kurie remiasi n-gramų žodynais ar sintakse [Tjong Kim Sang and Buchholz, 2000, Choueka, 1988, Smadja, 1993, Lin, 1998]. Segmentavimas pastoviaisiais junginiais naudoja junglumo reikšmių tarp šalia einančių žodžių kreivę kaip diskretųjį signalą, ir ribas, kurios skaido tekstą į žodžių sekas.

Segmentavimo pastoviaisiais junginiais labiausiai nuo kitų metodų skiriasi tuo, kad: (1) segmentavimas pastoviaisiais junginiais neskaido įterptųjų pastoviųjų junginių – imamas ilgiausias galimas duotame kontekste, tuo tarpu n-gramomis paremtas metodas negali tiesiogiai nustatyti ar pastovusis junginys yra įterptas kitame (pvz., *machine translation system*); (2) segmentavimas pastoviaisiais junginiais vyksta labai greitai, o sudėtingumas atitinka bigramų žodyno dydį, tuo tarpu n-gramomis pagrįstas būdas dažniausiai yra apribotas iki trijų žodžių ilgio pastoviųjų junginių ir yra žymiai didesnio sudėtingumo.

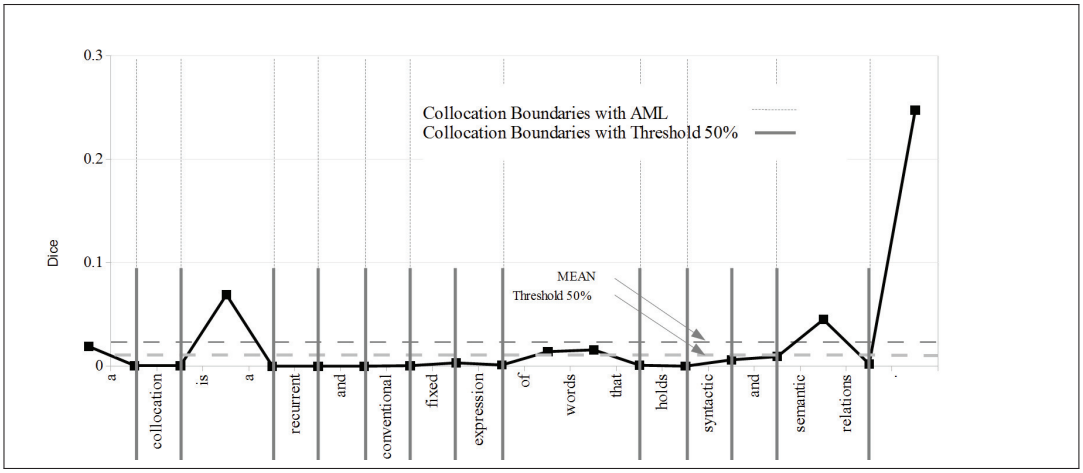

Figure 2.1: Sakinio segmentavimas pastoviaisiais junginiais *a collocation is a recurrent and conventional fixed expression of words that holds syntactic and semantic relations* [Xue et al., 2006].

Naujasis metodas naudoja unigramas ir bigramas junglumo stiprumo tarp šalia esančių žodžių nustatymui ir junglumo kreivės sudarymui. Tekstas yra matomas kaip besikeičianti

junglumo tarp šalia esančių žodžių kreivė (žiūr. pav. 2.1). Šis būdas yra naujas. Junglumo reikšmių kreivė yra naudojama pastoviųjų junginių segmentų ribų nustatymui. Pastoviojo junginio segmentas yra vienetas, kuris yra naudojamas kaip viena leksema ar unigrama n-graminiuose kalbos modeliuose.

Riba tarp dviejų šalia tekste esančių žodžių yra dedama kai junglumo reikšmė yra žemiau nustatyto slenksčio. Aš apibrėžiu slenkstį kaip atkarpą tarp sakinio junglumo reikšmių minimalios ir vidurkio reikšmių. Nulis atitinka minimalią junglumo reikšmę, o šimtas - sakinio junglumo reikšmių vidurkį. Taigi, slenkstis yra procentalė tarp sakinio junglumo minimalios ir vidurkio reikšmių. Tai leidžia taikyti tuos pačius nustatymus įvairiems junglumo vertinimo metodams.

Riba tarp dviejų šalia tekste esančių žodžių yra dedama kai junglumo reikšmė yra žemiau prieš tai ir po to esančių junglumo reikšmių vidurkį. Ši taisyklė gali būti taikoma kartu su slenksčiu ar be jo. Pastarasis tyrimas [Daudaravičius, 2012b] rodo, kad ši taisyklė leidžia pasiekti geriausius tekstų klasifikavimo rezultatus, tuo tarpu slenksčio naudojimas rezultatus blogina.

3 Segmentavimo pastoviaisiais junginiais taikymai

3.1 Automatinis daugiakalbis ES teisės dokumentų anotavimas EuroVoc deskriptoriais

Automatinis dokumentų anotavimas naudojant tezaurą yra naudingas norint nustatyti ryšius tarp panašių dokumentų. Daugiakalbiui automatiniam dokumentų anotavimui reikalinga suprasti kalbų skirtumus ir požymius, kurie gali būti naudingi tekstų klasifikavimui. Šis tyrimas pristato nuo kalbos nepriklausantį dokumentų anotavimo sistemą, kuri naudoja segmentavimo pastoviaisiais junginiais požymius.

Šiame tyrime naudoju tris labai skirtingas kalbas – anglų, lietuvių ir suomių. Šios kalbos skiriasi kaitybos ir žodžių dūrinių požiūriais. Suomių kalba yra labai kaitoma ir turi daug žodžių dūrinių. Lietuvių kalba yra labai kaitoma, tačiau žodžių dūriniai yra labai reti. Anglų kalba yra mažiausiai kaitoma ir naudoja labai mažai žodžių dūrinių.

Naudojant rankiniu būdu anotuotą daugiakalbį tekstyną *Acquis Communautaire 3.0 (AC)* [Steinberger et al., 2006] ir visus jame aptiktus deskriptorius, buvo pasiektas deskriptorių priskyrimo 5–8 % tikslumo pagerėjimas (t.y., nuo 48–54% pradinių) trims skirtingoms testuotoms kalboms (anglų, lietuvių ir suomių). Buvo pastebėta koreliacija tarp automatinio priskyrimo tikslumo ir dokumento ilgio. Šio tyrimo rezultatai atskleidė, kad skirtingas segmentavimo pastoviaisiais junginiais taikomas slenkstis duoda skirtingus požymius, kurie gali būti naudojami įvairiems kitiems tikslams. Nulinis slenkstis (t. y., jokio slenksčio) leido pasiekti geriausius tekstų klasifikavimo rezultatus.

3.2 Integravimas į statistinį mašininį vertimą

Kartu su *Universitat Politècnica de Catalunya* (UPC) ir *Barcelona Media Innovation Center* (BMIC), parengėme standartinį frazėmis grįstą satistinio mašininio vertimo sistemą ir dalyvavome *IWSLT 2010 MT* vertinimo veikloje. Naujajam segmentavimui pastoviaisiais junginiais buvo taikyti įvairūs statistiniai žodžių junglumo įverčiai, tokie kaip: Log–likelihood, T–score, Chi–squared, Dice, Mutual Information ir Gravity–Counts. Naujų segmentų naudojimas statistinio mašininio vertimo sistemoje leidžia praturtinti vertimo žodyną ir/arba glotnina esamas vertimo tikimybes. Mes sprendėme prancūzų-anglų *BTEC* užduotį. Mūsų pirminė ir kontrastuojanti sistemos buvo standartinės frazėmis grįstos

statistinio mašininio vertimo sistemos naudojant skirtingus segmentavimo pastoviaisiais junginiais rezultatus. Mes apjungėme standartinio frazėmis grįsto statistinio mašininio segmentavimą [Och and Ney, 2004] ir dvikalbiais segmentais, gautais taikant segmentavimą pastoviaisiais junginiais.

Vertimo rezultatai leido suskirstyti junglumo įverčius į tris grupes. Pirma, Log–likelihood, Chi–square ir T–score apjungia didelio dažnumo žodžius ir pastoviųjų junginių segmentai yra trumpi. Šie segmentai gerina statistinio mašininio vertimo sistemą įvedant naujus vertimo vienetus. Antra, Mutual Information ir Dice apjungia mažo dažnumo žodžius ir pastoviųjų junginių segmentai yra trumpi. Šie segmentai gerina statistinio vertimo rezultatus glotninant vertimo vienetų tikimybes. Trečia, Gravity–Counts apjungia didelio ir mažo dažnumo žodžius ir segmentai yra ilgi. Visgi, šiuo atveju statistinio mašininio vertimo rezultatai nebuvo pagerinti. Eksperimentiniai rezultatai pateikti prancūzų-anglų IWSLT 2010 automatiniam vertinimui buvo įvertina trečioje vietoje iš devynių dalyvių.

Mes atlikome tyrimą norėdami nustatyti ar segmentavimas pastoviaisiais junginiais duoda naudą tikimybių glotninimui ir naujų vertimo vienetų sudarymui. Nustatėme, kad segmentavimas naudojant Dice ir Mutual Information leidžia sėkmingai glotninti esamas vertimų tikimybes, o segmentavimas naudojant Chi–squared, Log–likelihood ar T–score leidžia sėkmingai sudaryti naujus vertimų vienetus.

3.3 Pritaikymas ACL ontologinių nuorodų tekstynui

ACL–2012 konfernecijos metu buvo organizuotas specialus seminaras *Rediscovering 50 Years of Discoveries* [Banchs, 2012], kad tumpam stabtelėti ir peržiūrėti praeitį bei įvertinti ateitį, įsitikinti, kad dabartinis palikimas išliks ilgam, ir kad ateities kartos galės saugiai kurti remdamiesi šios srities pionierių darbais. Pagrindinis šio seminaro tikslas buvo įvertinti kompiuterinės lingvistikos istoriją, evoliuciją ir ateitį. Organizatoriai ypatingai skyrė dėmesį pateiktiems tiriamiesiems darbams, kurie taikė natūralios kalbos apdorojimo ir tekstų kasinėjimo metodus ACL ontologinių nuorodų tekstynui (ACL Anthology Reference Corpus – ACL ARC), kuris yra viešai prieinamas ACL ARC projekto interneto tinklalapyje[1].

Terminijos suradimo metodai dažniausiai remiasi daugiažodžių vienetų suradimu naudojant n-gramas ir sintaksines fiksuotas sekas [Seretan, 2011]. Pagrindiniai šios užduoties

[1] `http://acl-arc.comp.nus.edu.sg/`

klausimai: (1) terminų ilgis – vienas–žodis, du–žodžiai, ar ilgesni terminai; (2) kalbos apdorojimo priemonių naudojimas, tokių kaip sintaksinis analizatorius; (3) kalbos aprėptis ir žodyno/terminijos kaita; (4) netolygus kasmetinis duomenų kiekio pasiskirstymas. Todėl, aš panaudojau segmentavimą pastoviaisiais junginiais norėdamas surasti pagrindines tyrimų metodų kryptis kompiuterinės lingvistikos bendruomenėje analizuojant terminijos varotseną. Pagrindiniai šio principo pricalumai yra: pirma, galimybė automatiškai nustatyti termino ilgį, ir antra, sistemos universalumas, nes nereikalinga naudoti specifinių kalbos įrankių(sintaksinių analizatorių ir kalbos dalių anotavimo įrankių).

Norėdamas įvertinti segmentavimo pastoviaisiais junginiais gebėjimą aptikti įvairius pastoviųjų junginių bruožus, aš suradau labiausiai reikšmingus pastoviųjų junginių segmentus ACL ARC tekstyne. Be to, pasinaudodamas į TF-IDF panašų rangavimą, aš suradau terminus, kurie yra susiję su įvairiomis natūralios kalbos analizės ir apdorojimo sritimis. Šių terminų pasiskirstymas ACL ARC tekstyne padėjo suprasti pagrindinius proveržius skirtingose tyrimų srityse metams bėgant. Iš kitos pusės, aš nesiekiau išsamios metodų, naudotų ACL ARC tekstyne, analizės, kuri yra sudėtinga ir ilga.

Analizė parodė, kad segmentavimas pastoviaisiais junginiais padeda surasti terminus dideliuose ir įvairiuose tekstynuose, kurie leidžia greitai ir paprastai analizuoti Kompiuterinės lingvistikos asociacijos istoriją. Rezultatai atskleidė, kad reikšmingi terminai iki 1986-ųjų yra susiję su formaliaisiais/taisykliniais tyrimų metodais. Pradedant nuo 1987-ųjų, su statistiniais metodais susiję terminai (pvz., *language model*, *similarity measure*, *text classification*) tampa labiau reikšmingi. 1990-aisiais, pastebimas didelis pokytis, kai terminai *Penn Treebank*, *Mutual Information*, *statistical parsing*, *bilingual corpus*, ir *dependency tree* tampa labai reikšmingi, rodantis, kad tyrimams naujose tyrimų kompiuterinės lingvistikos srityse didelę įtaką daro naujų kalbos išteklių sukūrimas. *Penn Treebank*, kuris yra reikšmingas tik trumpą laiką, yra naudojamas ligi šiolei. Pastarųjų metų reikšmingi terminai yra *BLEU score* ir *semantic role labeling*. Nors *machine translation* kaip terminas yra riekšmingas viso ACL ARC tekstyno atžvilgiu, tačiau jis nėra reikšmingas kokiu nors konkrečiu laiko momentu. Tai rodo, kad kai kurie terminai yra reikšmingi globaliai, bet nereikšmingi lokaliai.

Išvados

Pagrindinis mano siekis šioje disertacijoje yra pateikti paprastą ir efektyvų metodą, leidžiantį taikyti pastoviuosius junginius įvairiuose natūralios kalbos apdorojimo uždaviniuose. Segmentavimas pastoviaisiais junginiais leido surasti teksto vienetus, kurie yra sėkmingai naudojami terminijos ir žodynų sudarymo uždaviniuose, sudaryti efektyvius požymių vektorius teksto klasifikavimo uždaviniui, ir pagerinti statistinio mašininio vertimo mokymui reikalingo dvikalbio teksto lygiagretinimą. Pagrindinės mano disertacijos išvados ir rezultatai yra:

1. Pristatytas naujas pastoviųjų junginių apdorojimo metodas, kuris yra visiškai automatis ir pritaikomas įvairiuose naturalios kalbos apdorojimo uždaviniuose.

2. Teksto klasifikavimo greitis ir sudėtingumas yra tiesiogiai proporcingas naudojamo žodyno dydžiui. Segmentavimas pastoviaisiais junginiais leidžia pasiekti tokius pat ar geresnius tekstų klasifikavimo rezultatus nei naudojant bigramas, tačiau sugaištas laikas yra artimas kaip naudojant unigramas. Segmentavimo pastoviaisiais junginiais algoritmo sudėtingumas yra lygus bigramų, reikalingų nustatyti žodžių tarpusavio junglumo stiprumą, paieškai žodyne, ir yra tiesiogiai priklausantis nuo teksto dydžio. Tai leidžia segmentavimą pastoviaisiais junginiais sėkmingai taikyti praktikoje.

3. Tekstų klasifikavimui segmentavimas pastoviaisiais junginiais leidžia pasiekti geresnius rezultatus nei naudojant unigramas ar bigramas. ES teisės dokumentų klasifikavimo EuroVoc deskritpriais tikslumas buvo pagerintas nuo 48.30$\pm$0.48% iki 54.67$\pm$0.78%, nuo 51.41$\pm$0.56% iki 59.15$\pm$0.48% ir nuo 54.05$\pm$0.39% iki 58.87$\pm$0.36% atitinkamai anglų, lietuvių ir suomių kalboms.

4. Geriausi klasifikavimo naudojant segmentavimą pastoviaisiais junginiais rezultatai buvo pasiekti nenaudojant jokio segmentavimo slenksčio. Tai prieštarauja tradiciniams pastoviųjų junginių suradimo metodams, kuriems slenksčio taikymas yra pagrindinis "gerų pastoviųjų junginių" atrankos kriterijus. Tokiu būdu, segmentavimas pastoviaisiais junginiais atveria naujas pastoviųjų junginių apdorojimo galimybes.

5. Segmentavimas pastoviaisiais junginiais, iš esmės, yra nuo kalbos nepriklausantis, ir šis metodas taikytinas įvairioms kalboms, ypač turinčioms mažesnius kalbinius išteklius. Pastaruoju metu yra nesudėtinga sudaryti neanotuotus tekstynus, kurių užtenka kad galima būtų taikyti segmentavimą pastoviaisiais junginiais.

6. Segmentavimas pastoviaisiais junginiais leidžia automatiškai apčiuopti prasminius vienetus ir gali padėti termininijos vartosenos tyrimuose.

Introduction

"The (single) word is not privileged in terms of meaning. [...] Lexical items can be single words, compounds, multiword units, phrases, and even idioms" [Teubert, 2005]. Collocation is a well–known linguistic phenomenon which has a long history of research and use [Daudaravičius, 2012a]. Surprisingly, the term *collocation* itself has no standard definition, and this dissertation does not attempt to analyze the theoretical background of *what a collocation is*. There is a fundamental divide between those researchers who believe in collocation as a serious theoretical concept –

- "The meaning of a collocation is not a straightforward composition of the meanings of its parts. For example, the meaning of *red tape* is completely different from the meaning of its components." [Wermter and Hahn, 2004]

- "Collocation is an expression of two or more words that corresponds to some conventional way of saying things. Sometimes, the notion of collocation is defined in terms of syntax (by possible part-of-speech patterns) or in terms of semantics (requiring collocations to exhibit non-compositional meaning)." [Franz and Milch, 2002]

– and those who regard it as a convenient descriptive label without theoretical status.

- "Collocation is the occurrence of two or more words within a short space of each other in a text." [Frantzi and Ananiadou, 1996]

- "In its most general sense, a collocation is a habitual or lexicalised word combination." [Weeds et al., 2004]

- "A collocation is a recurrent sequence of words that co-occur together more than expected by chance in a given domain." [Gil and Dias, 2003]

Linguists study collocations from a lexical point of view. Both Melčuk [Melc'uk and Polguere, 1987] and the more recent computational linguists [Pecina, 2005] have recognized the need to define collocation phenomena from both the statistical and syntactical points of view.

Natural language processing community uses collocations as a key issue for tasks like natural language parsing (syntactic and semantic structure analysis) and generation (translation), as well as real-life applications such as machine translation, information extraction and retrieval. The general context of the term *collocation* is highly related to these concepts:

– **word** – the base unit of a language. Words are not simple phenomena. In general, a word is a partial or full representation of some concept in the world. While Indo-European languages write words using strings of individual characters, in Asian languages there is no clear definition of what a word is.

– **token** – an atomic element, usually a word in a natural language processing field. Tokenization is the process of splitting a text into a sequence of tokens. Simple words are easy to recognize, but there are words with no clear boundary. For instance, the token *don't* can be tokenized as: a single word | *don't* |, two words | *don* | *'t* | or | *do* | *n't* |, or three words | *don* | *'* | *t* |. There are no clear guidelines for how to tokenize a text into atomic units. Usually this decision is left to the engineers themselves. A recent study [He and Kayaalp, 2006] shows that the large number of freely available tokenizers produce distinct results (eleven out of thirteen are distinct).

– n**-gram** – a contiguous sequence of n words from a given text. Word n-grams are widely used in natural language modeling using HMM. They are also used to produce multidimensional vector spaces for text classification, machine translation, information retrieval, etc.

Collocation phenomena are simple to understand, yet hard to employ in real tasks. Little work has been done to simplify the use and application of collocations over the last few decades. [Smadja, 1991] introduced a general three-step method of collocation processing that presented 80% precision and 94% recall of the collocation extraction system. The system was used to compile a collocation dictionary of about 2,000 collocations – a lexicographic task. In general, the proposed algorithm allows collocations of two or more words in length to be extracted efficiently. While the results showed great success and a breakthrough in collocation extraction, this collocation extraction technique has never been applied to other tasks. The main obstacle to applying this approach was the complexity of the algorithm and the computational resources required. Even nowadays, applications that could make use of the implementation of such a system are barely possible in practice. Thus, for no good reason, collocations "were killed", and no new methods were introduced for a long time. Meanwhile, n-grams became the main strategy in natural language processing tasks: they are easy to use, simple to implement, and theoretically grounded. Collocations continued to be studied by enthusiasts. Beginning in the 1990s, computers became powerful enough to apply statistical methods (n-grams), but linguisti-

cally grounded, rule-based methods (collocations) remained too complex for the available resources. Thus, n-grams replaced collocations as the method of choice in NLP.

The importance of the collocation paradigm shift is raised in the most recent study on collocations, [Seretan, 2011]. The n-gram approach lacks the main property of language– *expression of meaning*, a descriptive property. How long should an n-gram be to capture meaning? The main approach is to keep the n-grams as long as possible without exhausting the technical resources or exceeding the time limit. This approach is practical, but poor in linguistic understanding. In practice, unigrams are applied when speed is required, and trigrams are used for higher quality results. The main practical issues for applying collocation are how far a collocation should be expanded, and what length it should be.

Higher order n-grams require large text corpora to extract reliable statistics. Such corpora exist for widely used languages such as English, Chinese, French, and German. The main issue here becomes less-resourced languages. While most studies are for the English language, there are many other languages to be processed. It is naive to believe that the methods applied to English can easily be applied to any other language. There are not many multilingual resources for comparative language processing. The largest multilingual corpus available is *The JRC–Acquis Multilingual Parallel Corpus*, compiled by the Joint Research Center of the European Commission. Until more corpora became available, there will not be many language processing tools, like parsers, for less-resourced languages.

The last issue in moving from n-grams to collocations is: can n-grams of different lengths be comparable? One recent approach is based on extracting bigram, trigram, quadgram, and five-gram collocations independently. However, while the main goal here is to recognize meaningful multiword collocations, there is no space left for single-word units, i.e., the status of single-word units (unigrams) is not clear, and unigrams are often omitted without justification.

In this dissertation I propose a new definition of the collocation: **A collocation is the longest non-overlapping combination of words in a text in which consecutive word pairs occur more often than by chance**. This definition extends the most common traditional definition of collocation (as *a recurrent sequence of words that co-occur together more than expected by chance*) with the requirement that the collocation processing technique be text-based, rather than n-gram filtering-based. In principle, this follows sentence segmentation issues: while it is hard to define *what a sentence is*, it is simple to recognize the start and end of a sentence without needing to analyze it first. My dissertation aims to

remedy the gap presented above by proposing a new type of segmentation.

The object of my dissertation are word collocations, which combine sequential words into meaningful (e.g., *machine translation*) or functional (e.g., *by the way*) lexical units in human languages.

The main goal of my dissertation is to introduce *collocation segmentation*, a collocation processing method that is: of low algorithmic complexity; not limited by collocation length; language-independent (only plain corpus); efficient; and applicable to many NLP tasks.

The main tasks that must be undertaken in order to accomplish this goal are:

- to summarize segmentation types, thereby presenting the main issues in defining segmentation units (text, discourse, topic, paragraph, sentence, word, etc.);
- to define an algorithmic procedure for collocation extraction, as well as to develop a general framework for collocation segmentation, including the selection of measures, features and thresholds that allow the collocation extraction system to be implemented;
- to find the optimal threshold for collocability, which measures the strength of relation between two words and defines the necessity to combine sequential words into one collocation, thus achieving the best text classification results;
- to study the influence of collocation segmentation on text classification for diverse languages (English, Lithuanian and Finnish), and to compare the results against unigram and bigram feature-based text classification;
- to study the influence of collocation segmentation, as compared to the standard unigram approach, on Statistical Machine Translation, by applying it to parallel text alignment;
- to apply collocation segmentation to a study of the main trends of research methods via the use of terminology in published articles.

The novelty and practical significance of the dissertation:

- A new, language-independent, and efficient collocation processing method is introduced, one which is applicable to many NLP tasks, enables a step back from the currently most widely adopted n-gram approach, and has acceptable computational complexity.

– A new approach for collocation processing, using a text as a curve of combinability values between two preceding words in the text, is investigated. The curve allows a new technique for collocation processing to be applied, based on the observation of the changing combinability values between two words. This approach avoids the necessity to define the threshold manually, which is obligatory in traditional collocation processing.

– A new framework for applying collocations to text classification is investigated. The similarities in classification performance in different languages (including English, Lithuanian and Finnish) show that collocation segmentation improves classification results and is applicable to diverse languages. Collocation segmentation does indeed produce the best results when compared to unigrams and bigrams.

– A new framework for applying collocations to Statistical Machine Translation is investigated. The analysis shows that collocation segmentation is especially beneficial to the Statistical Machine Translation system in the parallel text alignment step.

– A new framework for applying collocations to terminology processing in large corpora is proposed and investigated. This study has shown that collocation segmentation helps in the extraction of significant terms, and in the analysis of term yearly significance.

The structure of the dissertation. Chapter 2 presents an overview of various units and their main identification techniques. The discussion shows that unit identification is often based on setting boundaries. I apply this practice to the collocation segmentation method, which I introduce in Chapter 3. Then, in Chapter 4, I apply collocation segmentation and study three diverse tasks to show the value of the new proposed method. It must be mentioned that the integration of collocations into the Statistical Machine Translation system was done in cooperation with the Barcelona Media Innovation Center. Finally, the main conclusions are listed at the end of the dissertation.

6 Collocation Segmentation

Collocation segmentation is a new type of segmentation whose goal is to detect *fixed word sequences* and to segment a text into word sequences called collocation segments. I use the definition of a sequence in the notion of one or more. Thus, a collocation segment is a sequence of one or more consecutive words with high inter-combinability relations. A collocation segment can be of any length (even a single word) and the length is not defined in advance. This definition differs from other collocation definitions that are usually based on n-gram lists [Tjong Kim Sang and Buchholz, 2000, Choueka, 1988, Smadja, 1993]. Collocation segmentation is related to collocation extraction using syntactic rules [Lin, 1998]. The syntax-based approach allows the extraction of collocations that are easier to describe, and the process of collocation extraction is well-controlled. On the other hand, the syntax-based approach is not easily applied to languages with fewer resources. Collocation segmentation is based on a discrete signal of combinability values between two consecutive words, and boundaries that are used to chunk a sequence of words.

The main differences separating collocation segmentation from other methods are: (1) collocation segmentation does not break up nested collocations – it takes the longest one possible in a given context, while the n-gram list-based approach cannot detect if a collocation is nested in another one (e.g., *machine translation system*); (2) collocation segmentation is able to process long collocations quickly with the complexity of a bigram list size, while the n-gram list-based approach is usually limited to three-word collocations and has high processing complexity.

The new method employs unigram and bigram frequencies to evaluate combinability between two words and to build a combinability curve. A text is seen as a changing curve of Dice values between two adjacent words (see Figure 6.1). This approach is new. The curve of combinability values is used to detect the boundaries of collocation segments, which can be done using a threshold or by following certain rules. A collocation segment is a unit, which is used as a single token in n-gram language models.

A boundary can be set between two adjacent words in a text when the combinability value is lower than a certain threshold. I use a dynamic threshold which defines the range between the minimum and the average combinability values of a sentence. Zero equals the minimum combinability value and 100 equals the average value of the sentence. Thus, the threshold value is expressed as a percentage between the minimum and the average combinability values. If the threshold is set to zero, then no threshold filtering is used and no collocation segment boundaries are set using the threshold. This approach allows

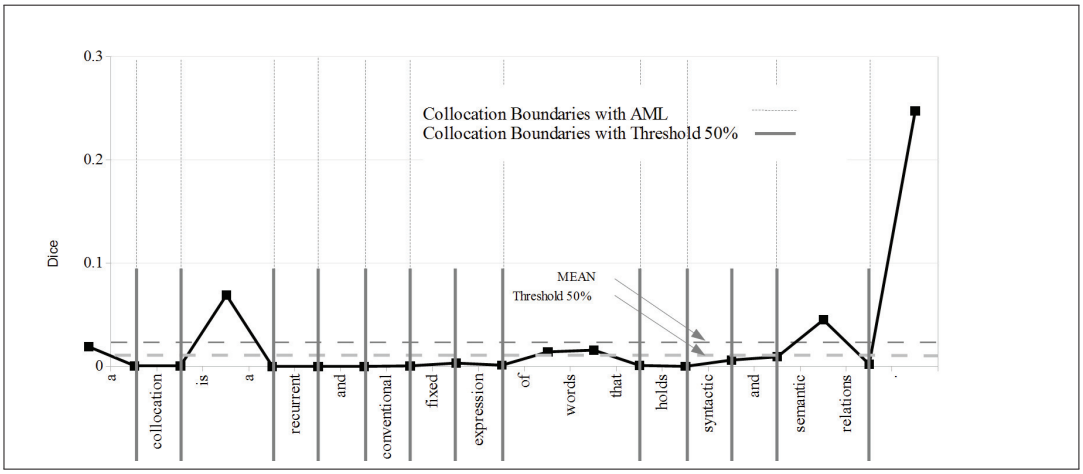

Figure 6.1: Collocation segmentation of the sentence *a collocation is a recurrent and conventional fixed expression of words that holds syntactic and semantic relations* [Xue et al., 2006].

comparable settings to be applied to different combinability measures.

A boundary is set between two adjacent words in a text when the Dice value is lower than the average of the preceding and following Dice values. The recent study of [Daudaravičius, 2012b] shows that Average Minimum Law is able to produce segmentation that gives the best text categorization results, while the threshold degrades them. On the other hand, AML can produce collocation segments when the combinability values between two adjacent words are very low (see Figure 6.1).

7 Applying Collocation Segmentation

7.1 Automatic multilingual annotation of EU legislation with EuroVoc descriptors

Automatic document annotation with a controlled conceptual thesaurus is useful for establishing precise links between similar documents. Automatic multilingual document annotation requires an understanding of the differences among the languages and features that are useful for the classification. This study presents a language-independent document annotation system based on features derived from collocation segmentation.

In this study I take three diverse languages – English, Lithuanian, and Finnish. These languages differ in inflection and agglutination aspects. The Finnish language is highly inflected and agglutinated, Lithuanian is highly inflected but undertakes no compounds, and English is the least inflected and agglutinated of the three.

Using the manually tagged multilingual corpus Acquis Communautaire 3.0 (AC) [Steinberger et al., 2006] and all of the descriptors found therein, I attained improvements in keyword assignment precision of 5–8 % (i.e., from 48–54% success) over the three diverse languages (English, Lithuanian and Finnish) tested. I found a high correlation between the precision of automatic assignment and document length. The results of this study show that different collocation segmentation thresholds introduce different features that can be used for various purposes. A zero threshold (meaning no threshold) allows the best text classification results to be achieved. This application-oriented technique shows whether collocations acquired by collocation segmentation are useful in document categorization tasks.

7.2 Integration into Statistical Machine Translation

Together with the Universitat Politècnica de Catalunya (UPC) and the Barcelona Media Innovation Center (BMIC), I have built a standard phrase-based SMT system enriched with novel segmentations, and participated in the IWSLT 2010 MT evaluation campaign. These novel segmentations are computed using statistical measures such as Log–likelihood, T–score, Chi–squared, Dice, Mutual Information or Gravity–Counts. Adding novel segmentations to an SMT system enriches the translation dictionary and/or smooths the existing translation probabilities. We participated in the French-to-English BTEC task, described

below. Our primary and contrastive systems were two standard phrase-based SMT systems enriched with different novel segmentations. We proposed to combine the standard phrase-based segmentation [Och and Ney, 2004] with a complementary bilingual segmentation learned from a statistical collocation segmentation technique. This statistical collocation segmentation uses measures such as the Dice score to estimate segments of words. The benefits of this procedure are twofold: (1) it extracts new translation units and (2) it smooths the probabilities of existing translation units.

The translation results allow the measures to be divided into three groups. First, Log–likelihood, Chi–square and T–score tend to combine high-frequency words and collocation segments that are very short. They improve the SMT system by adding new translation units. Second, Mutual Information and Dice tend to combine low-frequency words and collocation segments that are short. They improve the SMT system by smoothing the translation units. And third, Gravity–Counts tend to combine high- and low-frequency words and collocation segments that are long. However, in this case, the SMT system is not improved.

Thus, the road-map for improving the translation system is to introduce novel phrases containing either low-frequency or high-frequency words. This is a difficult method of improving translation quality. The experimental results were reported in the French-to-English IWSLT 2010 automatic evaluation, where our system was ranked third out of nine systems.

We performed an experiment to determine whether collocation segmentation benefits from smoothing the existing baseline phrases or introducing new phrases. We can conclude that segmentation with Dice or Mutual Information allows the existing baseline phrases to be smoothed, while segmentation with Chi–squared, Log–likelihood or T–score allows new phrases to be introduced.

7.3 Application to the ACL Anthology Reference Corpus

The ACL–2012 Special Workshop on Rediscovering 50 Years of Discoveries [Banchs, 2012] was organized to take a brief pause to review the past and project the future, to ensure that the current legacy will endure the indifference of time, and that future generations will be able to build securely on the foundations laid by the many pioneers of this discipline, in which language is clearly showing its reluctance to being tamed by mathematics.

The main objective of the workshop was to gather contributions about the history, the evolution and the future of research in computational linguistics. The organizers especially encouraged the submission of research projects that applied natural language processing and text mining techniques to the ACL Anthology Reference Corpus (ACL ARC), which is publicly available from the ACL ARC project website[1]. The main goal of this workshop was to investigate the efficacy of these techniques for computational history in general. A second goal was to use the ACL Anthology Reference Corpus [Bird et al., 2008] to answer specific questions about the history of computational linguistics. What is the path that the ACL has taken throughout its fifty-year history? What roles did various research topics play in the ACL's development? What have been the pivotal turning points?

Terminology extraction techniques are mainly based on the extraction of multiword units using n-grams and syntactic phrase patterns [Seretan, 2011]. The main challenges of this task are: (1) the length of the terms – one-word, two-word, or longer terms; (2) the use of linguistic processing tools such as parsers; (3) language coverage and vocabulary/terminology changes; (4) uneven annual redistribution of the data amount. Therefore, I applied collocation segmentation to study the main trends of research methods in the ACL community through the use of terminology. The main advantages of this approach are, first, the ability to automatically detect the length of the term, and second, the multilingualism of the system, because no linguistic processing (POS tagging or parsing) is required.

To evaluate the ability of collocation segmentation to handle different aspects of collocations, I extracted the most significant collocation segments in the ACL Anthology Reference Corpus. In addition, based on a ranking like that of $TF\text{-}IDF$, I extracted terms that are related to different phenomena of natural language analysis and processing. The distribution of these terms in the ACL ARC helps to understand the main breakthroughs of different research areas across the years. On the other hand, I did not intend to make a thorough study of the methods used by the ACL ARC, as such a task is complex and prohibitively extensive.

This study has shown that collocation segmentation can help extract terms from large and complex corpora, which speeds up research and simplifies the study of ACL history. The results show that the most significant terms prior to 1986 are related to formal/rule based research methods. Beginning in 1987, terms related to statistical methods (e.g., *language model*, *similarity measure*, *text classification*) become more important. In 1990, a

[1] `http://acl-arc.comp.nus.edu.sg/`

major turning point appears, when the terms *Penn Treebank*, *Mutual Information*, *statistical parsing*, *bilingual corpus*, and *dependency tree* become the most important, showing that research into new areas of computational linguistics is supported by the publication of new language resources. The *Penn Treebank*, which was only significant temporarily, is still used today. The most recent terms are *BLEU score* and *semantic role labeling*. While *machine translation* as a term is significant throughout the ACL ARC, it is not significant in any particular time period. This shows that some terms can be significant globally, but insignificant at a local level.

Conclusions

My aim in writing this dissertation was to provide a simple and efficient method for applying collocations to many NLP tasks. I have presented collocation segmentation, which allows the production of text units that may be used successfully in many terminology and lexicon extraction tasks, the building of efficient feature lists for text classification tasks, and the improvement of bilingual alignment for training Statistical Machine Translation systems.

The main outcomes of my dissertation are as follows:

1. A new collocation processing method was presented, one which is fully automatic and applicable to various NLP tasks, e.g., machine translation, text classification, lexicography, etc.

2. The speed and complexity of the classification task is directly proportional to the size of the dictionary used. Collocation segmentation achieves classification results that are equal to or better than those achieved using bigrams, while the processing time is close to that of unigrams. This is possible because the complexity of the collocation segmentation method is equal to the search in dictionary size of bigrams used for measuring the combinability of two consecutive words, and is linear to the length of the text. This makes collocation segmentation highly applicable in practice.

3. For the classification task, collocation segmentation outperforms other methods such as unigram and bigram. The precision of the categorization of EU legislation using EuroVoc was improved from 48.30±0.48% to 54.67±0.78%, from 51.41±0.56% to 59.15±0.48% and from 54.05±0.39% to 58.87±0.36% for the English, Lithuanian and Finnish languages respectively.

4. The best results in the classification task are achieved using collocation segmentation without any threshold applied. This contradicts the traditional view towards collocation extraction techniques, in which the threshold is the main criterion for filtering "good collocations". The collocation segmentation method thus opens new research areas in collocation processing.

5. Collocation segmentation, in general, is language-independent, and the method can be applied to many languages, especially those less resourced. Nowadays it is simple to acquire raw corpora that provide sufficient data for the performance of collocation segmentation.

6. Collocation segmentation is capable of handling meaningful units and can help in terminology studies.

Bibliography

Rafael E. Banchs, editor. *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*. Association for Computational Linguistics, Jeju Island, Korea, July 2012. URL http://www.aclweb.org/anthology/W12-32.

Steven Bird, Robert Dale, Bonnie J. Dorr, Bryan R. Gibson, Mark Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir R. Radev, and Yee Fan Tan. The acl anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In *LREC*. European Language Resources Association, 2008.

Yaacov Choueka. Looking for needles in a haystack or locating interesting collocational expressions in large textual databases. In *Proceedings of the RIAO Conference on User-Oriented Content-Based Text and Image Handling*, pages 609–624, March 1988.

Vidas Daudaravičius. Applying collocation segmentatio to the acl anthology reference corpus. In *Proceedings of Rediscovering 50 years of discoveries Workshop*, pages 10–18, Jeju, Rep. of Korea, July 2012a. Association for Computational Linguistics.

Vidas Daudaravičius. Automatic multilingual annotation of eu legislation with eurovoc descriptors. In *In Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'2012)*, pages 2142–2147, 2012b.

Katerina T. Frantzi and Sophia Ananiadou. Extracting nested collocations. In *COLING 1996: The 16th International Conference on Computational Linguistics*, 1996.

Alexander Franz and Brian Milch. Searching the web by voice. In *COLING 2002: The 17th International Conference on Computational Linguistics*, 2002.

Alexandre Gil and Gaël Dias. Using masks, suffix array-based data structures and multidimensional arrays to compute positional ngram statistics from corpora. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 25–32, Sapporo, Japan, July 2003. Association for Computational Linguistics. doi: 10.3115/1119282.1119286. URL http://www.aclweb.org/anthology/W03-1804.

Ying He and Mehmet Kayaalp. A comparison of 13 tokenizers on medline. Technical Report LHNCBC-TR-2006–003, The Lister Hill National Center for Biomedical Communications, 2006.

D. Lin. Extracting collocations from text corpora. In *First Workshop on Computational Terminology*, Montreal, 1998.

Igor A. Melc'uk and Alain Polguere. A formal lexicon in the meaning-text theory or (how to do lexica with words). *Computational Linguistics*, 13(3–4):261–275, July–December 1987. ISSN 0362-613X.

F.J. Och and H. Ney. The alignment template approach to statistical machine translation. *Computational linguistics*, 30(4):417–449, December 2004.

Pavel Pecina. An extensive empirical study of collocation extraction methods. In *Proceedings of the ACL Student Research Workshop*, pages 13–18, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P/P05/P05-2003`.

Violeta Seretan. *Syntax-Based Collocation Extraction*, volume 44 of *Text, Speech and Language Technology*. Springer, 2011. doi: 10.1007/978-94-007-0134-2_1. ISBN 978-94-007-0133-5.

Frank Smadja. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19: 143–177, 1993.

Frank A. Smadja. From n-grams to collocations: An evaluation of xtract. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pages 279–284, Berkeley, California, USA, June 1991. Association for Computational Linguistics. doi: 10.3115/981344.981380. URL `http://www.aclweb.org/anthology/P91-1036`.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiş, and Dániel Varga. The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, LREC'2006, pages 2142–2147, Genoa, Italy, May 2006.

Wolfgang Teubert. My version of corpus linguistics. *International Journal of Corpus Linguistics*, 10(1):1–13, 2005. doi: 10.1075/ijcl.10.1.01teu.

Erik F. Tjong Kim Sang and Sabine Buchholz. Introduction to the conll-2000 shared task: chunking. In *Proceedings of the 2nd workshop on Learning language in logic and the*

4th conference on Computational natural language learning - Volume 7, ConLL '00, pages 127–132, Stroudsburg, PA, USA, 2000. Association for Computational Linguistics.

Julie Weeds, David Weir, and Diana McCarthy. Characterising measures of lexical distributional similarity. In *Proceedings of Coling 2004*, pages 1015–1021, Geneva, Switzerland, Aug 23–Aug 27 2004. COLING.

Joachim Wermter and Udo Hahn. Collocation extraction based on modifiability statistics. In *Proceedings of Coling 2004*, pages 980–986, Geneva, Switzerland, Aug 23–Aug 27 2004. COLING.

Nianwen Xue, Jinying Chen, and Martha Palmer. Aligning features with sense distinction dimensions. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 921–928, Sydney, Australia, July 2006. Association for Computational Linguistics.

PUBLIKACIJOS DISERTACIJOS TEMA
(PUBLICATIONS ON THE TOPIC OF DISSERTATION)

KITUOSE MOKSLINĖS INFORMACIJOS INSTITUTO (ISI) DUOMENŲ BAZĖSE REFERUOJAMUOSE LEIDINIUOSE

1. **Daudaravičius, Vidas**. *The influence of collocation segmentation and top 10 items to keyword assignment performance*. Computational Linguistics and Intelligent Text Processing: 11th International Conference, CICLing 2010, Iasi, Romania, March 21-27, 2010. Springer. ISBN 9783642121159. p. 648-660.

KONFERENCIJŲ PRANEŠIMŲ MEDŽIAGOJE

2. Henríquez, Carlos A. Q.; Costa-jussà, Marta R.; **Daudaravičius, Vidas**; Banchs, Rafael E.; Mariño, B. José. *Using collocation segmentation to augment the phrase table*. WMT'10: Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR, Uppsala, Sweden, 15-16 July 2010. Association for Computational Linguistics. ISBN 9781932432718. p. 98-102.

3. **Daudaravičius, Vidas**. *Automatic multilingual annotation of EU legislation with Eurovoc descriptors*. In proceedings of 8th international conference on Language resources and evaluation, LREC 2012, 21-27 May, 2012, Istanbul, Turkey. European Language Resources Association. ISBN 9782951740877. p. 14-20.

4. **Vidas Daudaravičius**. *Applying Collocation Segmentation to the ACL Anthology Reference Corpus*. In Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries, Jeju Island, Korea, July 2012. Association for Computational Linguistics. ISBN 978-1-937284-29-9. p. 66–75.

KITUOSE PERIODINIUOSE LEIDINIUOSE

5. **Daudaravičius, Vidas**. *Automatic identification of lexical units*. Informatica: An International Journal of Computing and Informatics. Ljubljana: Slovensko društvo Informatika. ISSN 0350-5596. Vol. 34, no. 1 (2010), p. 85-91.

6. Costa-jussà, Marta R.; **Daudaravičius, Vidas**; Banchs, Rafael E.. *Using collocation segmentation to extract translation units in a phrase-based statistical machine translation system*. Procesamiento del Lenguaje Natural. Barcelona : Sociedad Espanola para el Procesamiento del Lenguaje Natural. ISSN 1135-5948. 2010, no. 45, p. 215-220.

KITOS PUBLIKACIJOS

7. Costa-jussà, Marta R.; **Daudaravičius, Vidas**; Banchs, Rafael E.. *Integration of statistical collocation segmentations in a phrase-based statistical machine translation system*. EAMT 2010: Proceedings of the 14th Annual Conference of the European Association for Machine Translation, 27-28 May 2010, Saint-Raphaël, France. European Association for Machine Translation. p. 1-8.

Vidas Daudaravičius (g. 1974) Vytauto Didžiojo universiteto Informatikos fakultete 2000 m. baigė informatikos bakalauro studijas, 2002 m. – taikomosios informatikos magistro studijas. 2008–2012 m. studijavo Vytauto Didžiojo universiteto fizinių mokslų srities informatikos krypties doktorantūroje. Nuo 2000 m. iki 2008 m. dirbo inžinieriumi–programuotoju Vytauto Didžiojo universiteto Kompiuterinės lingvistikos centre. Nuo 2003 m. dėsto Kompiuterinės lingvistikos kursą Vytauto Didžiojo universitete. Nuo 2011 m. dirba mokslo darbuotoju VTEX – mokslo publikavimo sprendimai – įmonėje.

Vidas Daudaravičius (b. 1974) acquired a BA degree in Informatics (2000) and an MA degree in Applied Informatics (2002) at the Faculty of Informatics, Vytautas Magnus University. In 2008–2012 he was a doctoral student at Vytautas Magnus University. From 2000 to 2008 he had been working as a engineer–programmer in the Centre of Computational Linguistics. Since 2011 has been working as an scientific researcher in VTEX – Solutions for Science Publishing – company.

Vidas DAUDARAVIČIUS

TEKSTO SKAIDYMAS PASTOVIŲJŲ JUNGINIŲ SEGMENTAIS

Daktaro disertacijos santrauka

Išleido ir spausdino – Vytauto Didžiojo universiteto leidykla
(S. Daukanto g. 27, LT-44249 Kaunas)
Užsakymo Nr. K12-179. Tiražas 20 egz. 2012 12 20.
Nemokamai.