

VILNIAUS UNIVERSITETAS
MATEMATIKOS IR INFORMATIKOS FAKULTETAS
PROGRAMŲ SISTEMŲ BAKALAURO STUDIJŲ PROGRAMA

Ilgų tekstų sugretinimas naudojant semantinius žodynus

Long text alignment using semantic dictionary

Bakalauro baigiamasis darbas

Atliko: Ieva Laukaitytė (parašas)

Darbo vadovas: doc. dr. Saulius Minkevičius (parašas)

Darbo recenzentas: dr. Haroldas Giedra (parašas)

Vilnius – 2020

Santrauka

Šį darbą sudaro teorinė ir eksperimentinė dalys. Teorinėje dalyje pirmiausiai yra aprašomi teksto gretinimo principai, populiariausi gretinimo algoritmai ir juos taikanti sistema OmegaT. Įvardijamos problemos išskylančios juos taikant ilgiems tekstams. Toliau yra aprašomos didžiosios ontologijos WordNet, FrameNet ir SUMO, jų naudojamos duomenų struktūros ir ryšiai tarp jų. Nagrinėjama kaip semantinę analizę atlieką CoreNLP sistema ir, WordNet ontologiją naudojanti, SEMAFOR sistema. Remiantis surinkta informacija yra pasiūlomas metodas, naudojantis semantinę informaciją apie žodžius, ilgo teksto gretinimui atlikti. Pasiūlyto metodo pagrindu sukuriamas prototipas. Eksperimentinėje dalyje atliekamas grožinės literatūros knygos gretinimas naudojant pasiūlytą metodą, stebima atlikto gretinimo kokybė. Gauti eksperimento rezultatai išnagrinėjami, pateikiamos išvados. Taip įrodoma, kad ontologijų naudojimas atliekant ilgo teksto gretinimą gerina gretinimo kokybę.

Raktiniai žodžiai: dvikalbių tekstų gretinimas, ontologijos, sakinių anotacijos sistemos, tekstynas, parafrazės.

Summary

This work consists of theoretical and experimental parts. The theoretical part describes the most popular text alignment algorithms and the systems that use them. It enlightens the problems that arise when applying them to long texts. The following describes the major ontologies WordNet, FrameNet, and SUMO. For semantic analysis systems SEMAFOR and CoreNLP are chosen to be compared. Based on the collected information method for long text alignment is proposed. A pilot version of alignment system is created and then tested. Aligning of a fictional book is performed. This process was being monitored and the quality of alignment was measured. The obtained experimental results are analyzed and presented. Based on the results of the experiment conclusions are made – semantic analysis improves long text alignment.

Keywords: bilingual text alignment, ontologies, semantic analysis systems, corpus, paraphrase.

TURINYS

IVADAS	6
1.GRETINIMO PRINCIPAI	8
1.1.Sakinių gretinimas	8
1.1.1.Leksiniai inkarai	8
1.1.2.Pagal žodžių skaičių sakinyje.....	8
1.1.3.Naudojant žodyną.....	8
1.1.4.Randant šablonus.....	9
1.2.Žodžių gretinimas	9
1.2.1.Remiantis statistiniais duomenimis.....	9
1.2.2.Remiantis lingvistika.....	10
2.TEKSTO GRETINIMO SISTEMOS IR JOSE TAIKOMŲ ALGORITMŲ APŽVALGA	11
2.1.Teksto gretinimo sistemos	11
2.1.1.OmegaT sistema	11
2.2.Kay ir Roscheisen algoritmas	12
2.2.1.Naudojamos duomenų struktūros.....	12
2.2.2.Algoritmo aprašas	13
2.2.3.Įeitis ir išeitis	13
2.3.Gale ir Church algoritmas.....	13
2.3.1.Algoritmo aprašas	14
2.3.2.Įeitis ir išeitis	14
3.ONTOLOGIJOS	15
3.1.WordNet.....	15
3.1.1.WordNet struktūra.....	15
3.1.2.Ryšiai tarp semantinių vienetų	16
3.2.FrameNet	16
3.2.1.FrameNet duomenų struktūros	17
3.2.2.Ryšiai tarp semantinių metaklasų.....	17
3.3.SUMO	18
3.3.1Struktūra	18
3.3.2.Ryšiai.....	19
4.SEMANTINĖS ANALIZĖS SISTEMOSE TAIKOMŲ ALGORITMŲ APŽVALGA	21
4.1.SEMAFOR algoritmas.....	21
4.1.1.Algoritmo aprašas	21
4.1.2.Įeitis ir išeitis	22

4.2.CoreNLP algoritmas	22
4.2.1.Algoritmo aprašas	22
4.2.2.Įeitis ir išeitis	23
5.SEMANTINĖS ANALIZĖS PRITAIKYMAS ILGO TEKSTO GREITINIME	24
5.1.Prototipas	25
5.1.1.Teksto filtravimo komponentas.....	27
5.1.2.Duomenų bazės komponentas	28
5.1.3.Statistinio sakinių greitinimo komponentas	30
5.1.4.Statistinio žodžių greitinimo komponentas	31
5.1.5.Semantinio žodžių greitinimo komponentas	31
5.1.6.Semantinio sakinių greitinimo komponentas	33
5.1.7.Žodžių vertimo komponentas	33
5.2.Eksperimentas	33
5.2.1.Eksperimento vykdymo aplinka.....	33
5.2.2.Pradiniai duomenys	34
5.2.3.Eksperimento eiga	34
5.2.4.Eksperimento rezultatai.....	35
REZULTATAI	37
IŠVADOS	38
ŠALTINIAI	39
SĄVOKŲ APIBRĖŽIMAI.....	41
PRIEDAI	42
1 priedas. Realizuoto prototipo vartotojo sąsajos ekranas	42
2 priedas. Žodžių semantinis greitinimas.....	43
3 priedas. Knygos greitinimas su OmegaT	45
4 priedas. Semantinis sakinių greitinimas.....	46

Ivadas

Teksto dalių sugretinimas – nurodymas vienos teksto dalies vertimą kitame tekste. Tai klasifikacijos uždavinys kai klasių skaičius yra labai didelis ir neapibrėžtas. Galutinis sugretinimo rezultatas yra lentelė pagal reikšmę sugrupuotų teksto dalių (vienos dalies apimtis nėra apibrėžta tiksliai, tai gali būti vienas žodis, sakiny, pastraipa). Teksto dalys gali būti tos pačios arba skirtingų kalbų, šiame darbe koncentruojamasi į skirtingų kalbų, anglų ir lietuvių, tekstų gretinimą. Daug skirtingų sistemų ir net mokslo šakų naudoja skaitmeninio teksto sugretinimo technologijas, jos pritaikomos vertimo kokybės įvertinime, vertimo klaidų paieškoje, informacijos išgavime, terminų nustatyme.

Darbe yra analizuojami jau egzistuojantys teksto gretinimo sprendimai: Kay – Roscheisen, Gale – Church algoritmai ir principai taikomi juose. Šie algoritmai arba jais paremtos modifikacijos yra naudojami populiariausių teksto gretinimo sistemų tokių kaip OmegaT, MateCat, Déjà Vu, GlobalSight, Across Language Server [CA16]. Jie remiasi panašiais principais – statistinę teksto analizę, vertimų atmintimi (angl. translation memory) ir loginėmis taisyklėmis. Abu nenaudoja išorinių informacijos šaltinių gretinimui atlikti ir kokybiškai gretina tik trumpos apimties tekstus ir griežtos struktūros dokumentus [ZAM14; YWZ10]. Kai naudojant šiuos algoritmus bandoma sugretinti ilgą tekstą (ilgu tekstas yra laikoma nuo 10 000 žodžių) jie nėra efektyvūs, gretinimo kokybė prasta ir geriausiu atveju tik 50% teisinga [Lau18]. Todėl šiame darbe bus koncentruojamasi į egzistuojančių gretinimo algoritmų pritaikymą ir patobulinimą ilgiems tekstams gretinti.

Semantinė analizė nusako kalbos supratimo procesą – kaip žmonės komunikuoja prasmėmis ir kontekstais. Ji identifikuoja teksto elementus, priskiria jiems reikšmę, gramatinę rolę. Žmogus informacijos apdorojimą ir prasmės paiešką daro nesąmoningai, pasitelkiant kultūrinį kontekstą, žinias apie kalbą, pasaulio suvokimą. Sistemos bando imituoti reikšmės supratimo procesą pasitelkiant semantinius žodynus. Semantinis žodynas, tai duomenų struktūra surišanti natūralios kalbos žodžius su ontologijos sąvokomis.

Šiuo darbu yra siekiama išsiaiškinti kaip panaudoti ontologijas gerinant tekstų gretinimą, koncentruojantis į ilgos apimties tekstus - pagrinde grožinės literatūros knygas. Grožinės knygos pasirinktos dėl savo sudėtingų ryšių tarp žodžių rečiau randamų kitokios literatūros veikaluose. Grožinėje literatūroje naudojama kalba yra vaizdinga, žodis ne visada vartojamas įprasta reikšme, randamos poetinės priemonės: frazeologizmai, sentimentai, ironija.

Darbo tikslas: sukurti ilgų tekstų gretinimo sistemos metodą ir jo pagrindu sukurti prototipą paremtą žodžių semantine analize, nustatyti ar ir kaip semantinė analizė gerina ilgų tekstų gretinimą.

Darbo tikslui pasiekti yra sprendžiami tokie uždaviniai:

1. Atlikti teksto gretinimo principų apžvalgą.
2. Atlikti egzistuojančių gretinimo sistemų ir jų taikomų algoritmų apžvalgą.
3. Atlikti didžiųjų ontologijų apžvalgą.
4. Apžvelgti egzistuojančias semantinės analizės sistemas.
5. Sukurti teksto gretinimo metodą naudojantį ontologijas.
6. Sukurti gretinimo sistemos prototipą atsižvelgiant į pasiūlytą metodą.
7. Ištestuoti kaip sukurtas metodas patobulina ilgo teksto gretinimą.

Remiantis knygos „Artificial Intelligence: A Modern Approach“ skyriais: „22.Communication“ ir „23.Probabilistic Language Processing“, interneto straipsniais ir gretinimo sistemų dokumentacijomis nagrinėjami gretinimo algoritmai, formuojamas naujas gretinimo metodas kuris bus naudojamas eksperimentinėje dalyje atliekant pasirinktos grožinės knygos gretinimą.

Darbą sudaro įvadas, penki pagrindiniai skyriai, rezultatų ir išvadų apibendrinimas, naudotos literatūros sąrašas ir priedai. Pirmuose darbo skyriuose apibrėžiamas ir detalizuojamas sprendžiamas uždavinys, aprašomi naudojami gretinimo principai ir algoritmai, pademonstruojamos problemos iškylančios juos naudojant ilgam tekstui gretinti. Trečiame ir ketvirtame skyriuose pasirinkti ir išanalizuoti kelios populiarios semantinės analizės sistemos ir jų algoritmai, kurie eksperimentinėje dalyje (penktame skyriuje) bus panaudoti atliekant ilgo teksto gretinimą.

1. Gretinimo principai

Teksto sugretinimo principai gali būti skirstomi pagal atliekamo gretinimo lygį. Tekstą sugretinti galima pagal sakinius ir žodžius [Ver00].

1.1. Sakinių gretinimas

Yra keturi sakinių gretinimo principai: leksiniai inkarai (angl. Lexical anchoring), pagal žodžių skaičių sakinyje (angl. Sentence length correlation), naudojant žodyną, randant šablonus (angl. Finding cognates)[ZAM14].

1.1.1. Leksiniai inkarai

Rankiniu būdu sugretinami sakiniai arba žodžiai kurie naudojami kaip tekstiniai inkarai padedantys išgauti gerą kitų sakinių gretinimą. Tokie gretinimą stebinčio žmogaus fiksuotos teksto elementų poros suskirsto visą tekstą į mažesnius vienas kitą atitinkančius segmentus [Ver00]. Teisingi sakinių/žodžių rinkiniai nėra šio principo tikslas, tai daugiau tarpinis procesas palengvintais kitus po jo einančius žingsnius.

1.1.2. Pagal žodžių skaičių sakinyje

Yra teigiama, kad sakinio ilgis yra tiesiogiai susijęs su to sakinio atitikmens kitame tekste ilgiu t.y. trumpo sakinio gretinimas bus trumpas, o ilgo – ilgas. Taikant šį principą naudojama tik vidinė tekstų informacija, nėra reikalingi jokie išoriniai duomenys ar žinios. Nėra daroma jokių išvadų apie sakinių struktūrą ir jos sąryšį su žodžiais jame [ZAM14]. Tai pat, manoma, kad skirtingom kalbom galima rasti bendrą koeficientą, kuris padeda nustatyti vieno sakinio vertimo ilgį kitoje kalboje [DQQ11]. Taikant tokį koeficientą, sakiniai yra grupuojami lyginant vieno su kito ilgį ir poziciją tekste.

1.1.3. Naudojant žodyną

Šis principas siūlo naudoti jau egzistuojantį sudarytą pradinį žodyną ir remiasi statistiniais duomenimis apie tų žodžių išsidėstymą tekste. Statistiniai duomenys yra labai naudingi nustatant nuo konteksto priklausančių žodžių vertimą ir reikšmę, o žodynai talpina informaciją apie žodžius

kurie tekste pasirodo tik vieną kartą ir todėl statistinių duomenų jiems gretinti nepakanka. Taigi, šis principas neatsižvelgia į žodžio kontekstą, todėl negali pagerinti daugiareikšmių žodžių gretinimo [HY97; YWZ10].

1.1.4. Randant šablonus

Šis principas remiasi tam tikrais kalbų panašumais. Randami simboliai, teksto elementai kurie nesikeičia išvertus į kitą kalbą arba išlieka panašūs [Ver00]. Dažnai nesikeičia datų, laiko, kainos formatavimas. Panašiais laikomi žodžiai kurių pirmos keturios raidės sutampa, tai gali būti veikėjo vardas, miestų pavadinimai, tarptautiniai žodžiai. Rasti šablonai naudojami panašiai kaip ir leksiniai inkarai – jie padalo tekstą į mažesnius segmentus. Daroma prielaida, kad gretinami šablonai bus gretinamuose sakiniuose. Skirtingai nuo leksinių inkarų, šablonai nėra patvirtinami žmogaus, šis procesas automatizuojamas.

1.2. Žodžių gretinimas

Dauguma sakinių gretinimo metodų bent jau dalinai naudoja žodžių gretinimo metodus. Tačiau teisingai atliktas žodžių sugretinimas nėra pagrindinis sakinių sugretinimo tikslas, o tik įrankis geriau atlikti sakinių sugretinimą, todėl jis atliekamas prastai ir yra randama daug neatitikimų.

Teksto elementai yra susieti vienas su kitu sudėtingais ryšiais - egzistuoja sudėtiniai žodžiai, frazeologizmai. Norint kokybiškai sugretinti žodžius reikia atsižvelgti į šiuos fenomenus. Neretai žodis neturi tiesioginio vertimo į kitą kalbą, kai kurie žodžiai neturi net panašaus atitikmens.

Žodžių gretinimo uždavinio rezultatas nevisada bus sugretintų žodžių poros, viename tekste vieną žodį atitinkanti reikšmė gali būti sudaryta iš kelių žodžių junginio kitame.

Žodžių gretinimo principus galima suskirstyti į dvi pagrindines grupes: remiantis statistika arba lingvistika (angl. Linguistic approaches) [ZAM14].

1.2.1. Remiantis statistiniais duomenimis

Statistiniai žodžių gretinimo metodai nenaudoja žinių apie kalbą t.y. jos gramatiką, leksikos taisykles, sakinių struktūrą. Žodžiai ir žodžių junginiai yra sugretinami atsižvelgiant į jų išsidėstymą tekste, pasikartojimo dažnį.

Naudojant šį principą susiduriama su problemom gretinant žodžius tekste pasikartojančius tik vieną kartą. Statistika paremti metodai sunkiai susitvarko su pavieniais, unikaliais ir mažai pasikartojančiais atvejais [XMY15].

1.2.2. Remiantis lingvistika

Naudojamos žinios apie kalbą: gramatiką, sakinių struktūros taisyklės. Šitas metodas labai brangus ir siauro naudojimo, nes žinios renkamos ir naudojamos tik su viena kalba, nėra naudingas nes labai siauras ir plačiai nepritaikomas [Ver00]. Šis principas remiasi tik morfologija ir neapima semantikos. Žodžiai gretinami pagal gramatiką ir kalbos dalis, o ne reikšmę.

2. Teksto gretinimo sistemos ir jose taikomų algoritmų apžvalga

Šiame skyriuje yra atlikta teksto gretinimo sistemų ir jų naudojamų algoritmų apžvalga.

2.1. Teksto gretinimo sistemos

Kompiuterinės sistemos atliekančios teksto sugretinimą joms duodamiems failams yra naudojamos informacijos išgavime, verčiant tekstus. Mašininės gretinimo sistemos vis dar nėra išstobulintos pakankamai, kad būtų galima atsisakyti žmogaus atliekant gretinimo procesą, jų išeitis vis dar turi būti tikrinama ir pertvarkoma vertėjo [TWC17]. Tai atliekama kontroliuojant suteikiamus įvairius įrankius, tokius kaip vertimų atmintis.

Vertimų atmintis - duomenų bazė sudaryta iš žodžių ir frazių, bei jų vertimų. Vertimų atmintis talpina prieš tai išverstus ir sugretintus bei sugrupuotus teksto fragmentus, dažniausiai sakinių ir pastraipų lygiu [Chr15]. Vertimų atmintis sudaroma sugretintą tekstą suskirsčius į segmentus ir juos sujungus vieną su kitu atitinkamai.

Populiariausios tekstų gretinimo sistemos: OmegaT, SmartCAT, MateCat, Déjà Vu, GlobalSight, Across Language Server [CA16]. Šiame darbe plačiau nagrinėti ir aprašyti pasirenkamas OmegaT sistemos algoritmas, nes ji yra vienintelė atviro kodo sistema naudojama profesionalių vertėjų.

2.1.1. OmegaT sistema

OmegaT - profesionali ir nemokama pagalbinė tekstą verčianti ir gretinanti sistema. Ji implementuoja tekstų gretinimo įrankį: naudotojas pasirenka du failus (sistema priima įvairius formatus) ir nurodo jų kalbas. Sugretinami tekstai yra pateikiami lentelės formatu arba kaip TMX failas. Gretinimo procesą sekantis žmogus turi galimybę keisti sugretintų sakinių poziciją pažymėdamas juos ir perkeldamas į kitą vietą.

Šis įrankis naudoja Gale ir Church algoritmą, taip pat suranda norimus šabloninius elementus: skaičius, datas ir pan. [WP14].

Šios sistemos teksto gretinimo kokybė priklauso nuo teksto apimties. Kol tekstą parduoda nedideliais segmentais, sudarytais iš 15 – 20 eilučių, klaidų skaičius ir dažnis yra nedidelis [Lau18].

OmegaT sistemai padavus nepravalytą (neparuoštą gretinimui) ir nepadalintą į segmentus knygą, jos gretinimas yra netikslus - sakiniai neatitinka vienas kito, tekste liko neišvalyti nereikalingi elementai (antraštės, puslapių numeriai, skyrių pavadinimai ir pan.). Net ir atfiltravus nereikalingus elementus gretinimo kokybė gaunasi labai prasta (žiūr. 3 priedą).

Tai pat, kuo ilgesnis gretinamas tekstas – tuo didesnė jo failo apimtis. Kuo didesnis failas, tuo ilgiau sistema užtrunka jį apdorodama. Kai OmegaIT paduodamo failo apimtis artėja iki 600 kB sistema pradeda striginėti [Lau18].

2.2. Kay ir Roscheisen algoritmas

Vienas iš populiariausių algoritmų kuris remiasi žodžių skaičiumi sakinyje. Šis algoritmas yra pritaikytas skirtingų kalbų tekstų gretinimui. Naudojama tik vidinė, iš teksto gaunama informacija [RKD06].

Algoritmas aprašomas pateikiant naudojamų duomenų struktūras, atliekamus žingsnius, įėjties ir išeities duomenis.

2.2.1. Naudojamos duomenų struktūros

Kay ir Roscheisen algoritmas naudoja keturias pagrindines duomenų struktūras:

1. Žodžio – sakinio indeksas (angl. Word-Sentence Index (toliau - WSI)). Paruošiamas abiem tekstams. Tai lentelė kuri savyje laiko kiekvieną tekste randamą žodį ir sakinį kuriame tas žodis randamas. Šiame kontekste žodis yra raidžių derinys, nebūtinai turintis reikšmę naudojamoje kalboje. Jei sakinyje žodis pasikartoja daugiau nei vieną kartą – sakinyje pasikartoja lentelėje daugiau nei vieną kartą.
2. Sugretinamų sakinių lentelė (angl. Alignable Sentence Table (toliau - AST)). Lentelė sudaryta iš galimai sugretinamų sakinių porų (po sakinį iš skirtingų tekstų).
3. Sugretinamų žodžių lentelė (angl. Word Alignment Table (toliau - WAT)). Žodžių porų sąrašas, sudarytas pagal jų išsidėstymą tekste. Lentelė talpina abiejų žodžių dažnius tekstuose, tai pat apskaičiuojamas žodžių panašumas.
4. Sugretintų sakinių lentelė SAT (angl. Sentence Alignment Table (toliau – SAT)). Lentelė kuri seka kiek kartų kiekviena sugretinamų sakinių pora buvo sugretinta. Dažniausiai pasitaikantis rezultatas bus pasirenkamas kaip galinis variantas.

2.2.2. Algoritmo aprašas

Procesas ciklinis. Kiekvieno ciklo pradžioje yra sukurama AST lentelė. Pradinė AST sudaryta pagal sakinių išsidėstymą tekstuose. Pirmo teksto pirmas sakiny yra suporuojamas su antro teksto pirmu sakiniu ir atitinkamai paskutinis su paskutiniu. Kuo sakiny yra arčiau teksto vidurio, tuo didesnis skaičius jo galimų porų [KR93]. Atliekamą procesą galima padalinti į tris žingsnius:

1. Sudaroma nauja WAT lentelė. Žodžių pora yra įtraukiama į WAT jei abu žodžiai yra randami sakinių poroje AST lentelė ir jų išsidėstymai abejose tekstuose yra pakankami panašūs.
2. Sukonstruojama SAT. Ji sudaroma remiantis jau sukonstruota WAT lentele, pirmenybė teikiama žodžių poroms pagal apskaičiuotą žodžių panašumą. Lentelėje gali būti keli to pačio sakinio gretinimo variantai.
3. Jei ciklas bus kartojamas, sudaroma nauja AST atsižvelgiant į seno ciklo SAT rezultatus. Jei ciklas nebus kartojamas, kaip rezultatas yra paduodama paskutinio ciklo metu gaunama SAT lentelė.

2.2.3. Įeitis ir išeitis

Įeitis – du, skirtingų kalbų tekstai.

Išeitis – paskutinio ciklo metu sukonstruota sugretintų sakinių lentelė (SAT).

2.3. Gale ir Church algoritmas

Gale ir Church algoritmas yra “vienas iš dažniausiai gretinimo sistemose naudojamų algoritmų”. Šis algoritmas “statistiškai apskaičiuoja lygiagrečių sakinių jungtis remdamasis ženklų skaičiumi sakiniuose” [RKD06]. Nenaudoja jokių išorinių žinių apie kalbą, gretinama remiantis tik turimais teksta ir jų vidine informacija.

Teigiama, kad sakinio ilgis yra susijęs su jo vertimo ilgiu. Trumpo sakinio vertimas bus trumpas, o ilgo – ilgas. Kiekvienai galimai porai yra paskaičiuojama tikimybė priklausomai nuo jų gretinamų sakinių ilgių (ilgis vertinamas pagal simbolių skaičių sakinyje) ir nuo ilgių pasiskirstymo tekste [GC93].

Algoritmas aprašomas pateikiant atliekamus žingsnius, įeities ir išeities duomenis.

2.3.1. Algoritmo aprašas

Atstumo matmuo (angl. distance measure) – matmuo naudojamas sakinių panašumui įvertinti. Apskaičiuojamas darant prielaidą, kad vienoje kalboje kiekvienas simbolis sukelia vertimą sudarančių simbolių pokytį.

Algoritmą sudaro du žingsniai, sugretinamos pastraipos ir tada sugretinami sakiniai esantys pastraipose [GC93]:

1. Sugretinamos pastraipos. Pastraipa yra laikomas teksto segmentas kurį sudaro daugiau nei 100 simbolių. Pastraipų sugretinimo tikslumas tikrinamas rankiniu būdu t.y. sugretinto teksto elementai patikrinami žmogaus.
2. Sugretinami sakiniai esantys sugretintose pastraipose. Algoritmas remiasi tuo, kad ilgesnių sakinių vertimas turėtų būti ilgesnis, o trumpesnių – trumpesnis. Yra sudaromos visų galimų gretinti sakinių poros ir kiekvienai porai pagal atstumo matmenį suskaičiuojama tikimybė t.y. yra sudaroma tikimybių lentelė.

2.3.2. Įeitis ir išeitis

Įeitis – skirtingų kalbų du tekstai.

Išeitis – pagal pastraipas suporuoti sakiniai. Dažniausiai vienas sakiny s pirmame tekste atitinka tik vieną sakinį antrame tekste, tačiau gali būti, kad vienam sakiniui bus priskiriami keli kiti sakiniai.

3. Ontologijos

Ontologija — aprašytos konkrečios srities (angl. domain) sąvokos bei jų ryšiai. Semantinių technologijų kontekste, ontologija suprantama, kaip universalus sąvokų, susietų ryšiais, aprašymo formatas leidžiantis vienodai suprasti ir atitinkamai komunikuoti tarp žmonių ir mašinų [Šil12]. Dažniausiai ontologijos nesurištos su natūralios kalbos žodžiais, bet yra išimčių, tokias ontologijos dar vadinamos semantinėmis leksinėmis duomenų bazėmis.

Bendroji ontologija – naudoja bendrąsias sąvokas, kurios tinka visoms sritim (objektas, savybė, sąsaja), todėl gali būti pagrindas daugeliui ontologijų, nėra konkrečios srities. Svarbiausia bendrųjų ontologijų savybė – jos sukuria būda konkrečių sričių ontologijoms apsikeisti informaciją (angl. semantic interoperability).

Ontologijos apibrėžia: sąvokas, ryšius tarp sąvokų, sąvokų tipus, aksiomas (taisykles) – dėsningumus tarp sąvokų. Jų struktūros yra skirtingos, bet visų primeną standartinę sandarą: klases, ryšius tarp klasių ir jų savybių [SPA07].

Šiame darbe aprašomos viena bendroji ontologija: SUMO, ir dvi leksinės duomenų bazės: WordNet, FrameNet. Ontologijos aprašomos išskiriant jų naudojamą sąvokas ir ryšius tarp jų.

3.1. WordNet

WordNet - leksinė duomenų bazė anglų kalba. Natūralios kalbos žodžiai yra surišti su unikalią loginę reikšmę reprezentuojančiais semantiniais vienetais.

Pagrindinis ryšys tarp WordNet žodžių yra sinonimija - žodžių ar žodžių junginių reikšmės tapatumas ir panašumas [STR04]. Šis leksikonas sujungia ne tik žodžius, o ir jų reikšmes. Sujungtų semantinių vienetų rezultatas yra naršyklę naviguojamas ontologinis reikšmių tinklas.

3.1.1. WordNet struktūra

WordNet-a sudaro 117 000 unikalių semantinių vienetų (angl. synsets) kurie yra tarpusavyje susiję skirtingais ryšiais. Semantinis vienetas - ontologijos klasė, kuri turi apibrėžimą natūraliaja kalba, bei ryšius su natūralios kalbos žodžiais [Fel12]. Semantinis vienetas reprezentuoja unikalią loginę prasmę kuri priskiriama su juo susietais žodžiais. Vienas žodis gali būti priskiriamas keliems semantiniams vienetais. Dauguma atvejų, su semantiniu vienetu surištas žodis turi pateiktus viena ar kelis jo naudojimo kontekste pavyzdinius sakinius.

Žodžiai su reikšmėmis yra surišti lemomis. Žodžio lema – jo pagrindinė forma, žodžio šaknis. Priskirdami semantiniui vienetui žodžių šaknis, o ne tiesiogiai juos pačius mes atsisakome skirtingų žodžio formų kurias jis gali įgyti – daugiskaitą, vienaskaitą, mažybinę ir pan. [Fel05]. Pavyzdžiui, lema {namas} bus surišta su žodžiais „namai“ ir „namas“ – abu žodžiai reiškia tą patį, skiriasi tik jų skaičius. Ši informacija yra išsaugoma atliekant sintaksės analizę, semantika nesidomi skirtingomis žodžių formomis.

3.1.2. Ryšiai tarp semantinių vienetų

Ryšiai tarp semantinių vienetų:

- Hiperonimia (angl. hyponymy) - sujungia abstrakčius semantinius vienetus su vis konkretesniais: {X} yra {Y} (X -hiponimas, Y - hiperonimas), {lova} yra {baldas}.
- Meronimija (angl. meronymy) - meronimas reiškia sudedamąją dalį arba narį: {X} yra {Y} dalis, {ratas} yra {automobilis} dalis.
- Ryšiai sudaryta pagal specifinę manierą kuri nusako ta patį įvykį: {komunikuoti}, {kalbėti}, {snibzdeti} – visi šitie veiksmažodžiai sujungti pagal semantinį vienetą {garsas}.
- Antonimai - žodžiai, turintys priešingas reikšmes, bet turintys bendrą turinio ypatybę, kurios skirtingas kiekis lemia antonimų poros susidarymą: {jaunas}, {senas} - bendroji ypatybė {amzius}.
- Morfologinis (angl. morphosemantic) - ryšys tarp tas pačias žodžio šaknis ir panašias prasmes turinčių žodžių: {stalas} ir {stalius}.
- Pertanimai (angl. pertainyms) - ryšys tarp būdvardžių ir daiktavardžių iš kurių būdvardis yra kilęs: {medis} ir {medinis}.

3.2. FrameNet

FrameNet - leksinė duomenų bazė anglų kalba. FrameNet sistema pagrįsta lingvistine teorija kuri teigia, kad loginę prasmę, kuri nusakoma natūralia žmogaus kalba, galima išreikšti ryšiais tarp natūralios kalbos žodžių ir ontologijos sąvokų.

FrameNet paremtas idėja, kad neįmanoma suprasti žodžio reikšmės be žinių susijusių su žodžio kontekstu. Žodis “pirkti” mums suprantamas todėl, kad suprantam kaip veikia pirkimo procesas, kas yra pardavėjas, mainai, pinigai, suprantam santykį tarp pirkėjo ir pardavėjo. Tai pagal

šita idėja kiekvienas žodis yra susietas su kitais žodžiais ir jų reikšmėmis - viena reikšmė nusakoma kitomis reikšmėmis [FJP03].

Kuriant FrameNet buvo nustatomos norimos aprašyti reikšmės, randami žodžiai atitinkantis tas reikšmes, sukuriama terminai reikšmei apibūdinti, gaunami ir anotuojami pavyzdiniai sakiniai. FrameNet sudarytas iš daugiau nei 13 000 reikšmių ir 200 000 rankiniu būdu anototų sakinių kurie susieti su semantinėmis FrameNet metaklasėmis.

3.2.1. FrameNet duomenų struktūros

Semantinė metaklasė - konkrečios situacijos realiaame pasaulyje reprezentacija, kurią mes galime išreikšti natūralios kalbos sakiniais. FrameNet semantinė metaklasė sudaryta iš metaklasės pavadinimo prasmės aprašymo, semantinės metaklasės elementų, lemu, pavyzdinių anototų sakinių kiekvienai lemai, ryšių su kitomis semantinėmis metaklasėmis [BFL98]:

- Semantinės metaklasės pavadinimo aprašas - pavadinimo prasmės paaiškinimas natūralia kalba.
- Semantinės klasės elementai (toliau FE) - semantinės rolės priskiriamos kiekvienai semantinei metaklasei. Šie elementai prideda daugiau informacijos semantinei sakinio struktūrai. Jie yra skirtingi kiekvienai semantinei metaklasei. FE skirstomi į pagrindinius (neatsiejamus nuo metaklasės reikšmės, jos visada egzistuoja) ir nepagrindinius (deskriptyvius).
- Leksiniai vienetai (angl. lexical units) – lemos (pagrindinė žodžio forma), su nurodyta kalbos dalimi. Viena lema gali būti priskiriama kelioms semantinėms metaklasėms. Kiekvienas leksinis vienetas turi viena ar kelis anotuotus sakiniu nusakančius jo naudojimo šabloną.
- Ryšiai su kitomis semantinėmis metaklasėmis.

3.2.2. Ryšiai tarp semantinių metaklasių

FrameNet išskiria šešis ryšius tarp semantinių metaklasių:

- Paveldimumas (angl. Inheritance) - viena semantinė metaklasė yra konkretesnė kitos metaklasės versija. Išvestinė metaklasė paveldi ir papildo bazinės klasės atributus (FE).
- Perspektyva (Perspectivized_in) - neutrali semantinė klasė sujungta su konkrečiu scenarijumi. Pavyzdžiui, {Commerce_transfer-goods} (lt. apsikeitimas resursais) yra sujungtas su {Commerce_sell} (lt. pardavimai), įgyjama pirkėjo perspektyva.
- Eiliškumas (angl. precedes) - nusako kompleksinio proceso eigą.

- Priežasties ir pasekmės ryšiai (angl. Causative_of and Inchoative_of) - nusako kada metaklasės naudojamos sąvokos yra susietos kaip priežastis ir pasekmė.
- Panaudojamumo (angl. using) - santykis tarp semantinių metaklasių, kai viena iš jų naudoja kitos elementus.
- Aiškinamieji (angl. See_also) - ryšys kurio formaliai apibrėžti negalime, bet jis palengvina vienos sąvokos suvokimą žiūrint į kitos sąvokos aprašymą.

3.3. SUMO

SUMO - aksiomatizuota bendroji ontologija. Kadangi ji naudoja bendrąsias sąvokas - tokias kaip klasė, esybė, laikas, objektas - ji nėra specializuota vienai sričiai (angl. domain) ir yra pagrindas konkretesnėms ontologijoms: ekonomikos, finansų, automobilių, muzikos, teisės, sporto, maisto it t.t. Iš viso ji į ją integruotos net trisdešimt dvis patvirtintos konkrečios ontologijos [Šev03].

Kitaip nei WordNet ar FrameNet, SUMO turi aksiomas, bet nėra surišta su natūralia kalba. SUMO neturi loginio suderinamumo (angl. logical inference) patvirtinimo sistemos. Loginio suderinamumo patvirtinimo sistema patikrintų ar aksiomos neprieštarauja viena kitai. Kadangi tai neatliekama SUMO aksiomose pasitaiko klaidų.

SUMO yra vienintelė formalioji ontologija kuri yra surišta su visu WordNet leksikonu [NP03a]. Tai buvo padaryta norint susieti SUMO sąvokas su natūralios kalbos žodžiais, taip pritaikant jį platesniam naudojimui.

3.3.1. Struktūra

Sumo ontologiją sudaro ~25,000 žinių bazės sąvokos (angl. KB Terms), susijusios įvairiais ryšiais, ir ~80,000 aksiomos (SUMO skelbiasi, kad yra viena iš labiausiai aksiomatizuotų formalųjų ontologijų).

Žinių bazės sąvokos – reikšmės reprezentacija žodžiais, prasmės paaiškinimas natūralia kalba. Ji dažniausiai sudaryta iš kelių žodžių abstrakčiai nusakančių formuojamą reikšmę [NP03b]. Ji gali būti paprasta – sudaryta iš vieno žodžio, pavyzdžiui {skystis} (angl. liquid) arba sudaryta iš kelių reikšmių intersekcijos ir suprantamas per kitų sąvokų ryšius su ja, pavyzdžiui {padidina tikimybę} (angl. increases likelihood).

SUMO aksiomos – SUO KIF (angl. standart upper ontology knowledge interchange format) kalba užrašyti taisyklės. SUO KIF - žinių mainų formatas yra kompiuterio kalba, sukurta tam, kad sistemos galėtų keistis žiniomis pagrįstų sistemų informacija ir ją pakartotinai naudoti.

Žinių bazės sąvokos {Paint} (lt. dažai) aksiomų pavyzdžiai:

```
(=>
  (instance ?PAINT Paint)
  (capability Coloring instrument ?PAINT))
```

(1)

(1) aksioma reiškia: jei objektas yra sąvokos **Paint** (lt. dažai) realizacija tai objektas gali būti panaudotas kaip instrumentas procese **Coloring** (lt. dažyti).

```
=>
  (and
    (instance ?PAINT Painting)
    (patient ?PAINT ?SURFACE)
    (resource ?PAINT ?STUFF))
  (exists (?OBJ)
    (and
      (surface ?SURFACE ?OBJ)
      (instance ?STUFF Paint))))
```

(2)

(2) aksioma reiškia: jei procesas (kurio adresas saugomas kintamajame ?PAINT) yra sąvokos **Painting** realizacija ir egzistuoja objektas ?SURFACE kuri dalyvauja procese ?PAINT ir egzistuoja objektas ?STUFF kuris yra proceso ?PAINT resursas, tai tada egzistuoja objektas ?OBJ ir šis objektas ?OBJ yra surištas su kitu objektu ?SURFACE ryšiu **surface** ir objektas ?STUFF turi tipą **Paint**. Natūralia kalba tai būtų galima užrašyti „Su dažais mes galime užtepti paviršių dažymo procese“.

3.3.2. Ryšiai

SUMO iš viso turi 164 skirtingus ryšius tarp reikšmių. Ryšį tarp SUMO terminų galima suprasti kaip predikatą. Predikatas – teiginys turintis parametrus (gražinantis dvi reikšmes: taip arba ne).

Sumo ryšiai naudojami formuojant aksiomas ir nusakyti teiginius apie modeliuojamą pasaulį.

Dauguma ryšių (~90%) pasikartoja mažiau nei 100 kartų. Dažniausiai pasikartojantys ryšiai sunumeruoti pagal dažnumą:

1. Realizacija (angl. instance) – konkretaus objektas, priklausantis klasei. Tai pat tai gali būti konkreti žinių bazės sąvoka. Pavyzdžiui, žinių bazės sąvoka {skystis} yra atitikmuo abstraktesnės sąvokos {fizinė_būseną}. Šis ryšys pasitaiko 30 882 kartus, dažniausiai pasitaikantis SUMO žinių bazėje.
2. Esantis (angl. located) – objektas randasi ties kitu objektu. Šis ryšys dažniausiai naudojamas enciklopedinėms žinioms pateikti (konkreto geografinio objekto surišimas su jo alokacija). Žinių bazėje juo surištos 20 600 žinių bazės sąvokos.
3. Narys (angl. member) – objektas yra objektu aibės narys (tas patas kaip meronimija) - {Y} dalis, {ratas} yra {automobilis} dalis. Šis ryšys pasitaiko 7 879 kartus.
4. Poklasis (angl. subclass) – žemesnė klasė kuri paveldi ryšius iš aukštesnių klasių. Paveldėtos klasės paprastai keičia tėvinės klasės funkcionalumą – prideda naujos informacijos. Dažnis: 7374.
5. Sritis (angl. domain) – susiejamos reikšmės, aprašo kitų ryšių kintamuosius. Dažnis: 2 992.

4. Semantinės analizės sistemose taikomų algoritmų apžvalga

Semantinės analizės sistemos, tai sistemos atliekančios įvairaus lygio semantinę analizę: anotavimą, sakinio skaldymą į kalbos dalis, koreferenciją, žodžių lemu radimas, pagrindinės žodžio formos radimą. Tokio sistemos plačiai naudojamos teksto tipo duomenų analizėje.

Šiame darbe plačiau nagrinėjamos ir aprašomos dvi populiariausios semantinę analizę atliekančios sistemos: SEMAFOR ir CoreNLP.

SEMAFOR – nekomercinė, moksliniams tyrimams sukurta, semantinės analizės sistema. SEMAFOR analize atlieka tik sakinių lygmenyje, todėl negali atlikti kokybiškos net trumpo teksto analizės. Ji neišsaugo informacijos tarp sakinių.

Ši sistema naudoja FrameNet ontologija – bando priskirti sakinio žodžius šios ontologijos metaklasėms [DSC+10].

Priešingai nei SEMAFOR, CoreNLP – semantinei analizei atlikti nenaudoja jau sudarytų leksikonų ar ontologijų. Ši sistema nenaudoja jokių išorinių resursų ir yra sudaryta iš kelių skirtingų, viena nuo kitos nepriklausančių, teksto analizės įrankių. Skirtingai nuo SEMAFOR, žodžio reikšmė čia nėra reprezentuojama duomenų struktūra panašia į semantinę metaklasę, o sluoksniuota informaciją „prisegama“ naudodama skirtingus anotatorius. CoreNLP išsaugo informaciją tarp sakinių - atlieka koreferenciją. Šis įrankis tai pat sugeba ištraukti žodžio lemas, nustatyti pagrindinę žodžio formą, normalizuoja data/laiką ir pan.. CoreNLP, skirtingai nei dauguma kitų semantinės analizės sistemų, gali atlikti skirtingą analizę keliomis kalbomis, netik anglų [MSB+14].

CoreNLP sudaryta iš aštuonių anotatorių, juos galima pasirinkti pagal reikmę, norimos teksto analizės lygį. Anotatorius – teksto analizės įrankis, asocijuojantis papildomą informaciją su konkrečia teksto dalimi.

4.1. SEMAFOR algoritmas

Naudojama FrameNet ontologija, bei tekstynas anotuotas FrameNet sąvokomis. Algoritmas aprašomas pateikiant atliekamus žingsnius, įėjties duomenis ir išeities duomenis.

4.1.1. Algoritmo aprašas

Procesas ciklinis – tekstas suskaidomas į sakinius ir kiekvienas sakinytis analizuojamas atskirai [CHE10]. Atliekamą procesą galima padalinti į keturis žingsnius:

1. Gramatinė analizė, sudaromas sakinio sintaksės medis.

2. Nagrinėjame sakinyje surandami žodžiai priklausantys semantinei metaklasei FrameNet ontologijoje. Rasti žodžiai yra pažymimi, bet nepriskiriami konkrečiai semantinei metaklasei.
3. Jei pažymėtas žodis yra priskiriamas kelioms semantinėms metaklasėms, atrenkama labiausiai tikėtina.
4. Semantinei klase nepriskirti žodžiai užpildo metaklasei priskiriamus semantinės klasės elementus (FE).

4.1.2. Įeitis ir išeitis

Įeitis – vienas sakiny. Galima paduoti tekstą, bet jis bus apdorojamas išskaidant juos į atskirus sakinius, pagal skiriamuosius sakinio ženklus. Neatliekama analizė tarp atskirų sakinių.

Išeitis – FrameNet ontologijos sąvokomis anotuotas sakiny.

4.2. CoreNLP algoritmas

CoreNLP teksto analizę atlieką naudodama kelis skirtingus įrankius – anotatorius. Analizės metu, priklausomai nuo norimos analizės, galima pasirinkti kokius anotatorius naudoti. Algoritmo apraše visi pateikiami žingsniai yra atliekami skirtingų įrankių ir yra apjungti bendros vartotojo sąsajos [MSB+14].

Algoritmas aprašomas pateikiant atliekamus žingsnius, įeities ir išeities duomenis.

4.2.1. Algoritmo aprašas

Anotacijos objektas – tekstas kurį anotuojame. Kiekvienas analizės anotatorius prideda informacijos prie anotacijos objekto.

Kiekvienam anotatoriui yra paduodamas tas pats anotacijos objektas. Atliekamą procesą galima padalinti į vienas po kito atliekamus šešis žingsnius:

1. Tekstas suskirstomas į žodžius ir frazes.
2. Tekstas dalomas į sakinius. Tai atliekama surandant skyrybos ženklus.
3. Lemų generavimas – išskiriama bazinė žodžio reikšmė.
4. Randami ir pažymimi vietovardžiai, organizacijos, žmonių vardai. Sprendžiamas klasifikavimo uždavinys – žodis gali būti priskiriamas vienai iš trijų pagrindinių klasių (PERSON, ORGANIZATION, LOCATION) (lt. ŽMOGUS, ORGANIZACIJA, VIETA).

5. Gramatikos ir sintaksės analizė – sudaromas sintaksės medis naudojant gramatines taisykles. Jis parodo kokios kategorijos priklauso sakinio žodžiai: veiksmažodis, daiktavardis,rieveiksmis, būdvardis ir dalelytės. Apskaičiuojama tikimybė kiekvienai galimai kalbos daliai ir priskiriama labiausiai tikėtina.
6. Koreferencija (angl. coreference) – nuoroda į žodį kuris turi tą pačią reikšmę kitame sakinyje, bet skirtingą leksinę išraišką, pvz. „Ieva pakėlė obuolį.“ ir „Ji įsidėjo obuolį į kišenę.“ – jei šie sakiniai tekste eina vienas po kito, žodžiai „Ieva“ ir „Ji“ reprezentuoja tą patį objektą ir turi vienodą reikšmę.

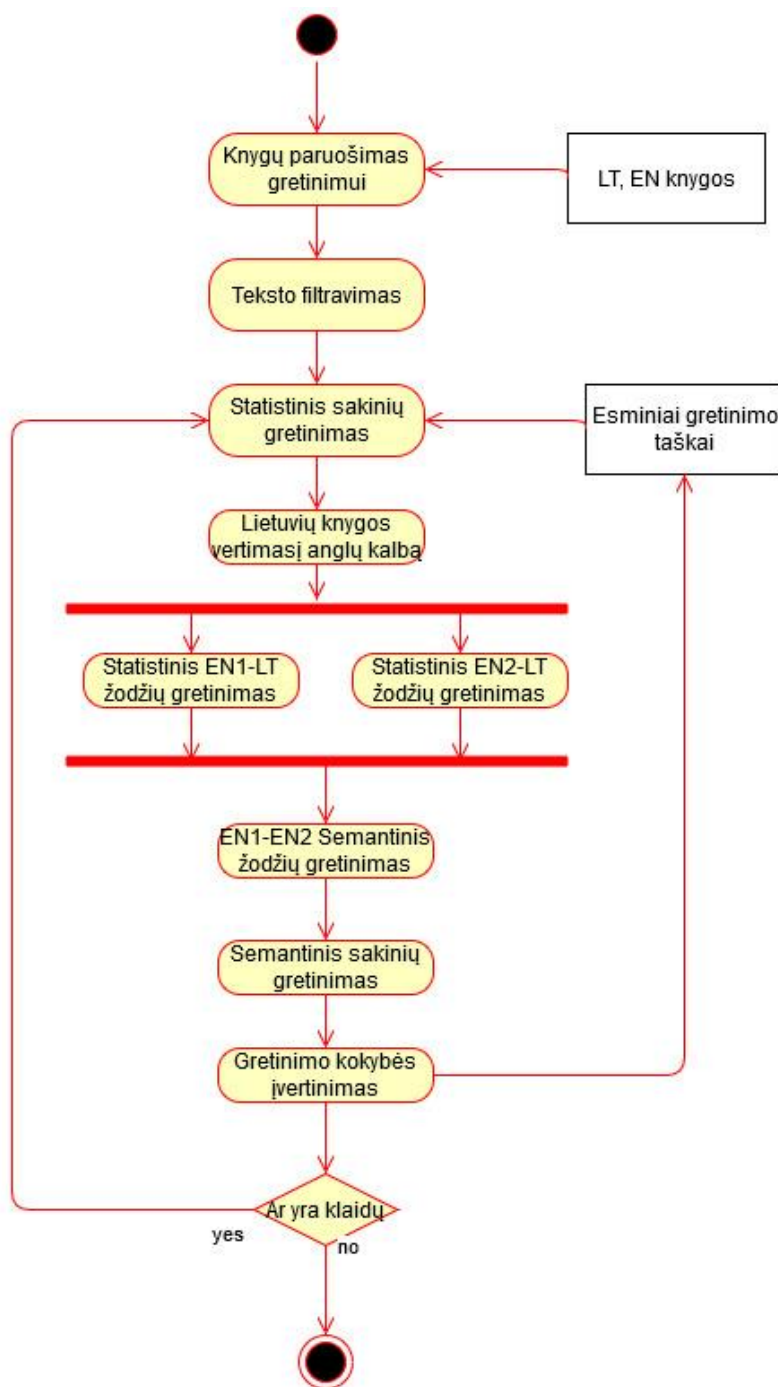
4.2.2. Įeitis ir išeitis

Įeitis – tekstas.

Išeitis – į kalbos dalis suskirstytas sakiny su išskirtais tarp - žodiniais ryšiais. Galima pasirinkti tarp XML failo ar paprasto teksto.

5. Semantinės analizės pritaikymas ilgo teksto gretinime

Atsižvelgiant į ankstesniuose skyriuose surinktą informaciją, pasiūlomas naujas gretinimo metodas. Metodo aprašymas pateikiamas naudojant UML veiklų diagramą (žiūr. 1 pav.) ir veiklų lentelę (žiūr. 1 lent.).



1 pav. Siūlomas ilgų tekstų gretinimo metodas

1 lentelė. Metodo veiklų aprašymas

Veiklos pavadinimas	Veiklos aprašymas
Knygų paruošimas gretinimui	Surinktos knygos, nepriklausomai nuo jų formato, konvertuojamos į txt failus.
Teksto filtravimas	Nereikalingų ilgo teksto elementų pašalinimas naudojant reguliarias išraiškas.
Statistinis sakinių gretinimas	Statistinių metodų naudojimas sakinių gretinimui gauti.
Lietuvių knygos vertimas į anglų kalbą	Automatinės vertinimo sistemos naudojimas lietuvių kalbos knygos vertimui į anglų kalbą.
Statistinis EN1 – LT žodžių gretinimas	Statistinių metodų naudojimas žodžių gretinimui gauti.
Statistinis EN2 – LT žodžių gretinimas	Statistinių metodų naudojimas žodžių gretinimui gauti.
EN1 – EN2 Semantinis žodžių gretinimas	Ontologijų naudojimas žodžių gretinimui gauti.
Semantinis sakinių gretinimas	Atlikto žodžių semantinio gretinimo rezultatų naudojimas atlikti sakinių semantiniui gretinimui.
Gretinimo kokybės įvertinimas	Atlikto gretinimo kokybės įvertinimas pagal pasirinktas metrikas naudojant auksinį gretinimo standartą.

Šio eksperimento tikslas yra sukurti eksperimentinį sprendimą, gretinimo sistemos prototipą paremtą pasiūlytu metodu, ir ištestuoti kaip sukurtas sprendimas leidžia pagerinti ilgo teksto gretinimą.

5.1. Prototipas

Siekiant nustatyti ar ir kaip semantinė analizė pagerina ilgų tekstų gretinimą sukurtas eksperimentinis sprendimas (žiūr. 2 pav.) gretinimui atlikti.

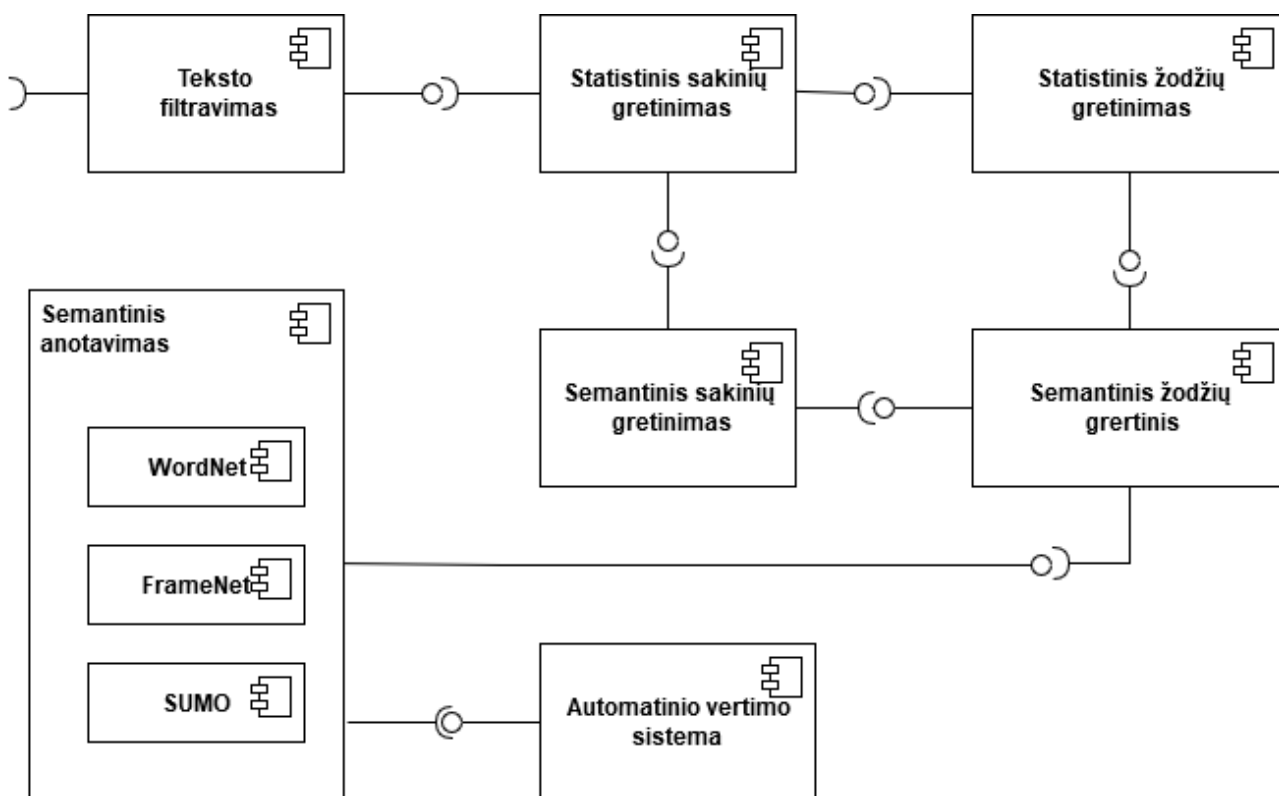
Šį sprendimą sudaro komponentai leidžiantys išbandyti skirtingus gretinimo metodus: naudojant ir nenaudojant semantinės analizės.

Nepriklausomai nuo gretinimo metodo visada atliekamas teksto paruošimas gretinimo procesui – tam naudojamas teksto filtravimo komponentas. Padavus nepravalytą knygą, jos gretinimas bus netikslus dėl tekste pasitaikančių nereikalingų elementų: antraščių, išnašų, vertėjo komentarų, skyrių ir puslapių numeracijos. Norint pasiekti gera gretinimo kokybę šie elementai turi būti pašalinami.

Pašalinus negretinamus teksto elementus atliekamas statistinis gretinimas. Skyriuje „Teksto gretinimo sistemos ir jose taikomų algoritmų apžvalga“ yra trumpai apibūdinti pagrindiniai teksto gretinimo algoritmai, remiantis jais atliekamas statistinis sakinių ir žodžių gretinimas.

Eksperimento metu semantiniai analizėi atlikti yra naudojamos visos trys skyriuje „Ontologijos“ aptartos ontologijos. Jos yra naudojamos atskirai ir kartu bandant nustatyti kuri yra naudingiausia pasirinktam uždaviniui atlikti.

Kadangi ontologijos yra surištos tik su anglų kalbos natūraliais žodžiais, tarp gretinimo ir semantinės analizės procesų reikia atlikti žodžių vertimą.



2 pav. Komponentų diagrama

Sukurto prototipo komponentai:

- Teksto filtravimo komponentas.
- Duomenų bazės komponentas.
- Sakinių statistinio gretinimo komponentas.
- Statistinio žodžių gretinimo komponentas.
- Semantinio sakinių gretinimo komponentas.
- Semantinio žodžių gretinimo komponentas.
- Teksto vertimo komponentas.

5.1.1. Teksto filtravimo komponentas

Teksto gretinimo algoritmai yra pritaikyti trumpam tekstui gretinti, todėl neatsižvelgia į problemą išskylančią gretinant ilgus tekstus: knygos turi nereikalingus teksto elementus kuriuos reikia pašalinti. Nereikalingi ir nesugretinami elementai: vertėjo komentarai, eilutės sudarytos tik iš skaičių, puslapio numeriai, užsilikę formatavimo elementai [Lau18]. Teksto filtravimo komponentas sprendžia iškilusią problemą atliekant nereikalingų elementų radimą ir pašalinimą. Tekstas yra pravalomas naudojant reguliarias išraiškas. Taip gaunamas švaresnis ir lengviau gretinamas knygos variantas. Netinkamu teksto elementu yra laikoma: antraštės, skyrių numeracija, vertėjo pastabos ir išnašos.

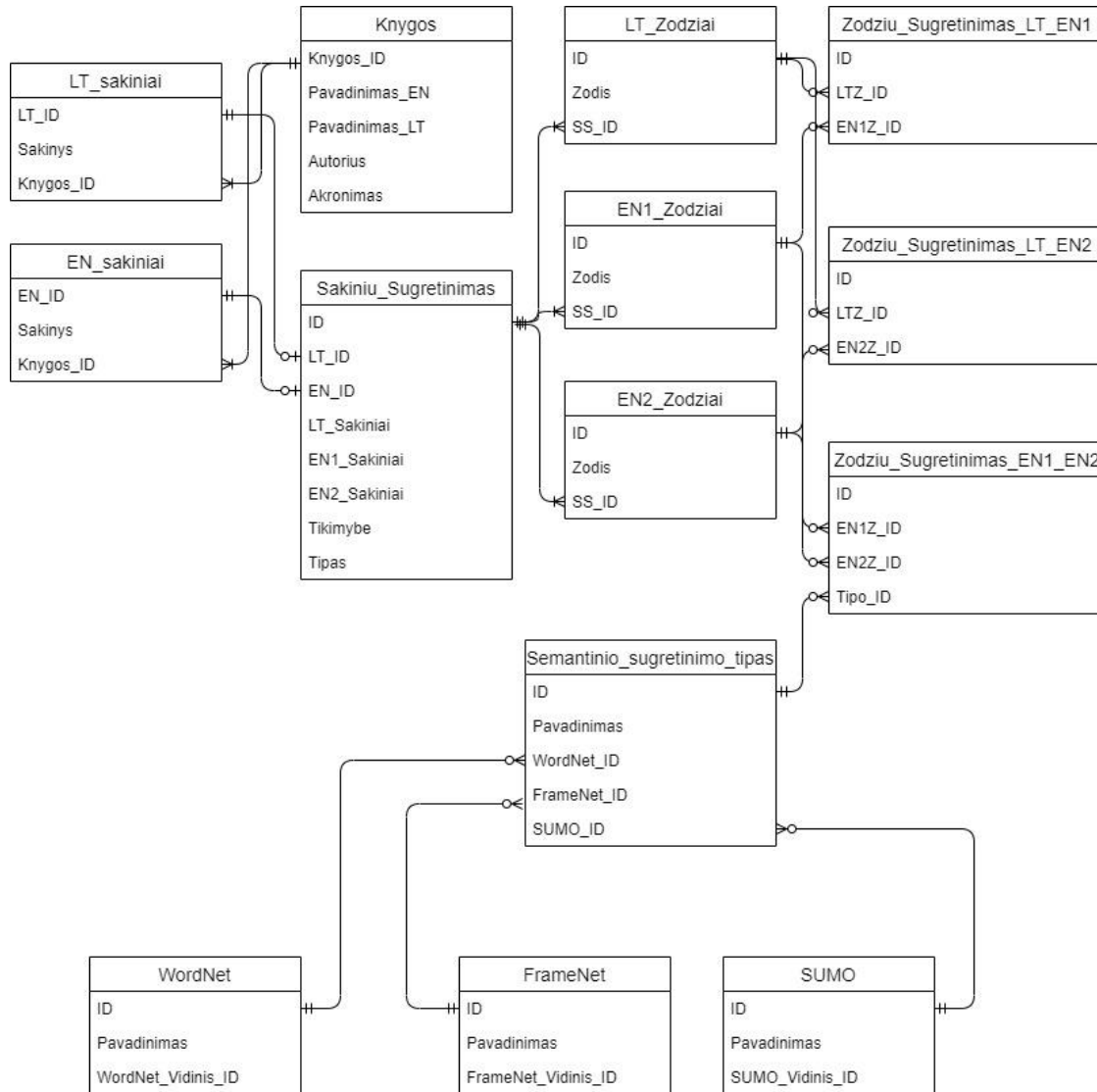
Teksto netinkamų elementų radimui naudojamos šios reguliarios išraiškos:

- $^{\wedge} * (d)\{1,3\} [\backslash d] * \$$ - Eilučių, sudarytų tik iš skaičių, šalinimas.
- $(p\{L\})(d)\{1,3\} [\backslash d] * r\backslash n$ - Puslapio numerių, prikabinėtų prie teksto eilučių, šalinimas.
- $(d)\{1,3\} [\backslash d] * \$$ - Surandami visi skaičiai. Tai naudinga norint atsikratyti visų skaitmenų, galimai prarandant net tekstui priklausančius skaičius.
- $^{\wedge} *, + \$$ - Vertėjo komentarų radimui.
- $^{\wedge} [\backslash d] * [\dots] \{10,30\} [\backslash d] * \$$ - Konkretiems pasikartojantiems elementams, kaip knygos pavadinimui, rasti.
- $(d)(p\{L\}) \quad \backslash 1 \backslash 2$ - Tarpelio tarp skaičiaus ir raidžių įterpimui.

5.1.2. Duomenų bazės komponentas

Duomenų bazė eksperimentiniame sprendime saugo įvairius gretinimo proceso komponentus: sugretintus tekstus sakinių lygmenyje, sugretintus žodžius, žodžių vertimus, semantinį žodžių gretinimą, papildomą informaciją apie gretinamus tekstus.

Pateikta reliacinė schema (žiūr. 3 pav.) parodo duomenų bazės struktūrą.



3 pav. Duomenų bazės reliacinė schema

Lentelės:

- LT_sakiniai(LT_ID, Sakinys, Knygos_ID) – lentelė sauganti lietuvių kalbos teksto fragmentus suskaidytus pagal skyrybos ženklus. LT_ID yra automatiškai generuojama id, o Knygos_ID , tai nuoroda į knygą iš kurios yra sakiny.
- EN_sakiniai(EN_ID, Sakinys, Knygos_ID) – lentelė sauganti anglų kalbos teksto fragmentus suskaidytus pagal skyrybos ženklus. EN_ID yra automatiškai generuojama id, o Knygos_ID , tai nuoroda į knygą iš kurios yra sakiny.
- Knygos(Knygos_ID, Pavadinimas_EN, Pavadinimas_LT, Autorius, Akronimas) – informacija apie gretinamas knygas. Akronimas – knygos sutrumpintas pavadinimas. Nors šio eksperimento metu gretinama tik viena knyga, reikia atsižvelgti į tai, kad gali būti gretinamos kelios – pritaikyti prototipą tekstyno kūrimui (angl. corpus).
- Sakiniu_Sugretinimas(ID, LT_ID, EN_ID, LT_Sakiniai, EN1_Sakiniai, EN2_Sakiniai, Tikimybė, Tipas) – lentelė surišanti LT_sakiniai ir EN_sakiniai sakinius. EN1_Sakiniai - originalas anglų kalba. EN2_Sakiniai – LT_Sakiniai vertimas į anglų kalbą. Tikimybė – sugretinimo poros įvertinimas. Tipas – pora gali būti patvirtinta žmogaus (inkaras), arba sugretinta automatiškai.
- LT_zodziai(ID, Zodis, SS_ID) – LT_sakiniai frazės suskaldytos į žodžius. Zodis – žodžiai sudarantys lietuvių kalbos teksto sakinius. SS_ID – nuoroda į Sakinių_Sugretinimas.
- EN1_zodziai(ID, Zodis, SS_ID) – EN1_sakiniai frazės suskaldytos į žodžius. Zodis – žodžiai sudarantys anglų kalbos originalaus teksto sakinius. SS_ID – nuoroda į Sakinių_Sugretinimas.
- EN2_zodziai(ID, Zodis, SS_ID) - EN2_Sakiniai frazės suskaldytos į žodžius. Zodis – žodžiai sudarantys versto lietuvių kalbos teksto sakinius. SS_ID – nuoroda į Sakinių_Sugretinimas.
- Zodziu_Sugretinimas_LT_EN1(ID, LTZ_ID, EN1Z_ID) – statistinio žodžių gretinimo rezultatai. Lietuvių kalbos teksto žodžių gretinimas su originalaus anglų kalbos tekstu.
- Zodziu_Sugretinimas_LT_EN2(ID, LTZ_ID, EN2Z_ID) – statistinio žodžių gretinimo rezultatai. Lietuvių kalbos teksto žodžių gretinimas su verstų sakinių žodžiais.
- Zodziu_Sugretinimas_EN1_EN2(ID, EN1Z_ID, EN2Z_ID, Tipo_ID) – semantinio žodžių gretinimo rezultatai. Lietuvių kalbos vertimo ir anglų kalbos originalo sugretintų žodžių poros.
- Semantinio_sugretinimo_tipas(ID, Pavadinimas, WordNet_ID, FrameNet_ID, SUMO_ID) – saugo sugretintų žodžių tipą, tai yra bendrąją reikšmę. Ji susideda iš skirtingų ontologijų

gražinamų reikšmių. WordNet_ID, SUMO_ID, FrameNet_ID – nuorodos į to tipo reikšmes atitinkamose ontologijose.

- WordNet(ID, Pavadinimas, WordNet_Vidinis_ID) – lentelė talpinanti informaciją apie konkrečią reikšmę, susiejanti ją su WordNet duomenų bazės vidiniu id.
- FrameNet(ID, Pavadinimas, FrameNet_Vidinis_ID) – lentelė talpinanti informaciją apie konkrečią reikšmę, susiejanti ją su FrameNet duomenų bazės vidiniu id.
- SUMO(ID, Pavadinimas, SUMO_Vidinis_ID) – lentelė talpinanti informaciją apie konkrečią reikšmę, susiejanti ją su SUMO duomenų bazės vidiniu id.

5.1.3. Statistinio sakinių gretinimo komponentas

Šis įrankis paremtas statistiniais sakinių gretinimo metodais. Programai paduodami du txt failai - dvi tos pačios knygos versijos (skirtingos kalbos). Gretinimui atlikti naudojama statistika - sakinių poros surandamos remiantis panašumo matmeniu sudaromoje tikimybių lentelėje. Šio komponento išeitis - po viena ar kelis sugrupuoti sakiniai arba sakinių fragmentai (žiūr. 3 priedą).

Sakinių gretinimui naudojamo algoritmo žingsniai:

1. Tekstai skaidomi pagal skiriamuosius ženklus - {.,:;-?!} į smulkius teksto fragmentus.
2. Pagal teksto fragmento poziciją tekste jam priskiriami galimi gretinimo variantai iš kito teksto. Tai daroma lyginant teksto apimtis ir remiantis idėja, kad ilgesnių sakinių vertimas turėtų būti ilgesnis, o trumpesnių – trumpesnis.
3. Yra sudaromos visų galimų gretinti sakinių poros ir kiekvienai porai apskaičiuojama tikimybė:

$$P(e_y, l_x) = \frac{\sum_{i=1}^N k_i \cdot f_i(e, l)}{\sum_{x=1}^L \sum_{i=1}^N k_i \cdot f_i(e, l)} \quad (1)$$

Formulės (1) parametras e_y – anglų kalbos sakiny (kurio indeksas tekstyne yra y), o l_x yra lietuvių kalbos sakiny (kurio indeksas x). $f_i()$ – sakinių panašumo funkcijos (angl. feature functions). k_i – sakinių panašumo funkcijos koeficientai. N – panašumo formulių skaičius.

Funkcijos naudojamos sakinių panašumui nusakyti:

- f_1 - pagal žodžių skaičių sakinyje.
- f_2 – ar abu sakiniai pasibaigia vienodais skyrybos ženklais - {?, !}
- f_3 – ar abu sakiniai pasibaigia tašku.

- f_4 – ar abu sakiniai prasideda ir baigiasi šnekamosios kalbos ženklais, randami dialogai.
 - f_5 – bendras kiekis surastų frazių naudojant žodyną.
 - f_6 – bendras kiekis surastų bendrų žodžių žodyne.
4. Atrenkamos geriausios sugretintų tekstų fragmentų poros, atliekama optimizacija.

5.1.4. Statistinio žodžių gretinimo komponentas

Veikia labai panašiai kaip ir sakinių gretinimo komponentas tik vietoj pilnų nesugretintų tekstų, įeities duomenys yra sakinių gretinimo komponento sugretintų sakinių poros.

Algoritmas – ciklinis. Įvardinti žingsniai atliekami kiekvienai sugretintų sakinių porai:

1. Imama viena sugretintų tekstų fragmentų pora ir gretinami žodžiai joje.
2. Abu teksto fragmentai suskaidomi pagal tarpus.
3. Yra sudaroma visų galimų sugretintų žodžių porų tikimybių lentelė, apskaičiuojama tikimybė kiekvienai žodžių porai:

$$P(e_y, l_x) = \frac{\sum_{i=1}^N k_i \cdot g_i(e, l)}{\sum_{x=1}^L \sum_{i=1}^N k_i \cdot g_i(e, l)} \quad (2)$$

Formulės (2) parametras e_y – anglų kalbos žodis (kurio indeksas sakinyje yra y), o l_x yra lietuvių kalbos žodis (kurio indeksas x). $g_i()$ – žodžių panašumo funkcijos. k_i – žodžių panašumo funkcijos koeficientai. N – panašumo formulių skaičius.

Funkcijos naudojamos žodžių panašumui nusakyti:

- g_1 – abu žodžiai prasideda iš didžiosios raidės.
 - g_2 – abu žodžiai vien tik didžiosios raidės.
 - g_3 – žodžių pora surasta žodyne.
 - g_4 – žodžių pora sudaro frazę, kuri rasta žodyne.
4. Optimizacija, atrenkamas geriausias variantas.

5.1.5. Semantinio žodžių gretinimo komponentas

Šis komponentas sudarytas iš ankstesniame skyriuje „Ontologijos“ išvardintų technologijų: WordNet, SUMO ir FrameNet.

Bandant pritaikyti ontologijas atlikti semantinę analizę teksto gretinimo sistemoje reikia atsižvelgti į iškylančias problemas:

- Semantinio žodyno daugiareikšmiškumą - žodis gali turėti kelias reikšmes kurios priklauso nuo konteksto kuriame žodis yra naudojamas.
- Aksiomų skaičius yra labai mažas, jų nepakanka nusakyti net mažai daliai egzistuojančių ryšių tarp reikšmių.
- SUMO neturi loginio suderinamumo (angl. logical inference), aksiomoms patvirtinti, sistemos. Pasitaiko logikos klaidų kurios gali pakenkti galutiniam rezultatui.
- Yra kalbos barjeras, išsamios ontologijos yra sukurtos ir pririštos tik prie anglų kalbos. Norint atlikti semantinę analizę kitų kalbų tekstams, juos reiki versti.

Kalbos barjeras pašalinamas naudojant vertimo sistemą – gretinami ne skirtingų kalbų žodžiai, o du anglų kalbos žodžiai (vienas originalas, o kitas vertimas iš lietuvių į anglų).

Po statistinio sakinių gretinimo yra gaunama lentelė surišanti teksto fragmentus, jų atitikmenis lietuvių kalbos tekste ir išverstą tekstą (žiūr. 4 priedą). Semantinis žodžių lyginimas atliekamas tarp originalo ir lietuvių vertimo į anglų. Kiekvienas sugretinto sakinio fragmentų žodis yra lyginamas vienas su kitu ir atitinkamai yra įvertinamas jų reikšmių ryšio stiprumas.

Ontologijos priskiria žodį reikšmei t.y. priskiria žodžiui reikšmės indeksą ir lyginant ontologijas mes lyginam atitinkamų jų indeksus. Žodžiai gali būti priskirti tai pačiai reikšmei arba jų reikšmės gali būti susietos ryšiu (žiūr. 2 priedą).

Jei semantinei analizei naudojamos visos ontologijos yra gaunami trys įvertinimai, tada galimai sugretinamų žodžių klasifikacijos lyginamos ir atitinkamai kiek panašių klasifikacijų jie turi, jų ryšio stiprumas įvertinamas nuo 0 iki 3 (0 – jei nesutampa visos klasifikacijos, 1 – jei sutampa viena ir t.t.). Jei naudojamos dvi ontologijos ryšio stiprumas vertinamas nuo 0 iki 2, o jei tik viena, įvertinama arba 0 arba 1.

Vertėjo patvirtinta žodžių kombinacija, ryšio įvertinimą pakėlus virš 3, žodžių pora tampa leksiniu inkaru – esminiu gretinimo tašku.

Šio gretinimo rezultatas – sugerintos pagal reikšmę žodžių poros (žiūr. 1 priedą).

5.1.6. Semantinio sakinių gretinimo komponentas

Semantinis sugretinimas, atvirkščiai nuo statistinio, pirmiausiai vykdomas žodžių lygmenyje. Po to, remiantis prielaida, kad vienos reikšmės sakiniai sudaryti iš vienodų reikšmių sumos t.y. žodžių, galima patikrinti jau esamus sakinių sugretinimus.

Semantinis sakinių gretinimas gali būti puikus būdas pataisyti statistinį sakinių gretinimą. Jei po statistinio gretinimo yra keli vienodai ar panašiai tikėtini gretinimo variantai, semantinis gretinimas gali patikslinti ir išrinkti tinkamą sakinių porą paskaičiuodamas kiek panašių reikšmių žodžių yra kiekvienoje galimoje poroje.

5.1.7. Žodžių vertimo komponentas

Ontologijos yra susietos tik su anglų kalbos žodynu, todėl norint pritaikyti jas skirtingų kalbų tekstų gretinimui reikia tuos žodžius išversti į anglų kalbą.

Vertimas atliekamas prieš statistinį gretinimą, nes norima išlaikyti semantinę sakinių informaciją, jei vertimas būtų atliekamas žodžiams, jis būtų netikslus ir prarastu kontekstą.

Tekstas verčiamas suskaidžius jį į mažus, po 10 sakinių, segmentus.

5.2. Eksperimentas

Atliktas eksperimentas aprašomas pateikiant vykdymo aplinką, jam atlikti naudojamus duomenis, eksperimento eigą ir gautus rezultatus.

5.2.1. Eksperimento vykdymo aplinka

Kompiuterio, kuriame buvo vykdomas eksperimentas, specifikacijos:

- Procesorius — AMD RYZEN 3600 (6 branduoliai, 12 loginių procesorių)
- Sisteminis diskas — SSD 240GB
- Operatyvioji atmintis — 32GB
- Operacinė sistema — Windows 10 EDU

5.2.2. Pradiniai duomenys

Gretinimui buvo naudojamas ilgos apimties tekstas – grožinės literatūros knyga Artūro Konano Doilio “Baskervilių šuo”. Atliekamas gretinimas tarp dviejų knygos versijų – tarp originalo anglų kalba ir vertimo lietuvių kalba (žiūr. 2 lentelę).

Norint po gretinimo patikrinti rezultatus buvo sudarytas auksinis standartas – žmogaus peržiūrėtas ir patvirtintas knygų gretinimas. Auksinis standartas turi du variantus: sakinių ir žodžių gretinimus. Jis buvo sudarytas knygą skaidant į mažus segmentus, gretinant naudojant sistemą OmegaT ir koreguotas rankomis. Šis standartas buvo naudojamas sistemos sugretinto teksto kokybės vertinimui.

2 lentelė. Eksperimentui naudojamų tekstų apimtys

	LT	EN
Sakinių skaičius	4065	4108
Žodžių skaičius	43355	59962
Simbolių skaičius	280575	310645

5.2.3. Eksperimento eiga

Pirmiausiai duomenys buvo paruošiami gretinimui. Tam naudojamas teksto filtravimo komponentas: su reguliariom išraiškom randami ir pašalinami negretinami teksto elementai. Tai buvo atliekama abiejų kalbų knygoms.

Eksperimentas buvo sudarytas iš dviejų dalių:

1. Tiriamas statistinio gretinimo efektyvumas gretinant skirtingos apimties tekstus. Randama iki kokios apimties tekstas bus gretinamas kokybiškai. Pradedama nuo mažos apimties segmento – keliasdešimt originalios knygos sakinių, toliau imamas kitas, didesnės apimties segmentas. Kiekvienam segmentui rankiniu būdu parenkamas atitinkamas vertimo knygos segmentas. Procesas vis kartojamas, gautas gretinimas lyginamas su sudarytu auksiniu standartu.
2. Antroje eksperimento dalyje, lyginamas pilnos, nepadalintos knygos gretinimas naudojant skirtingas ontologijas ir jų kombinacijas. Bandoma nustatyti ar semantinės analizės taikymas gretinimo procese pagerins ilgo teksto gretinimo kokybę. Gauti rezultatai lyginami su pirmos dalies duomenimis.

5.2.4. Eksperimento rezultatai

Atlikto gretinimo uždavinio tikslumą vertinti pasirinktos trys metrikos:

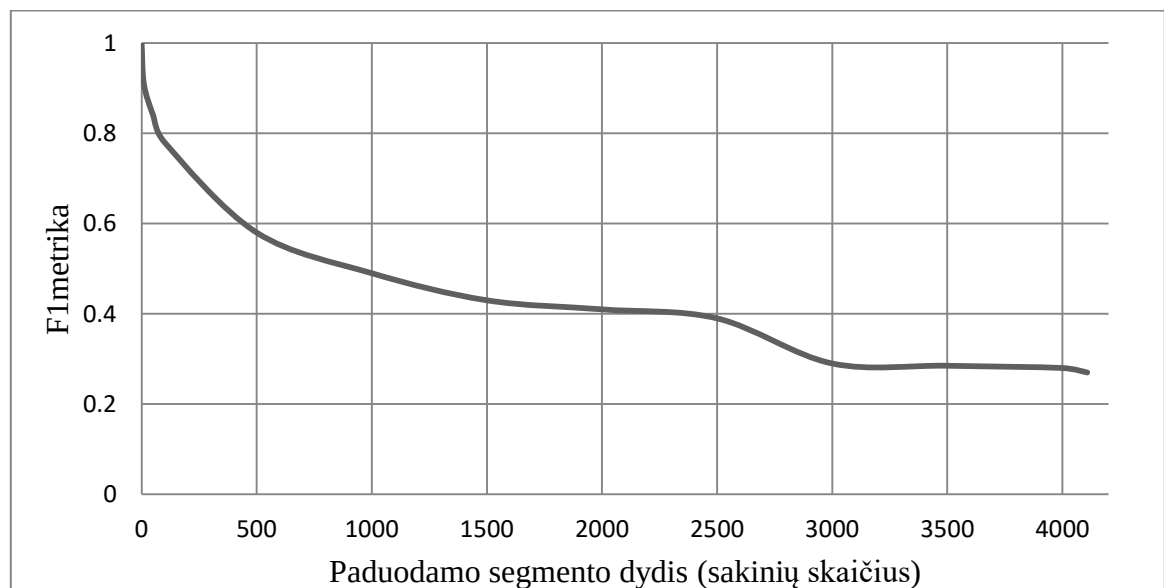
- tikslumas (angl. precision) - kaip teisingai klasifikuotų objektų ir nustatytų objektų santykis.
- atpažintų arba klasifikuotų objektų kiekis (angl. recall) - kaip teisingai klasifikuotų objektų ir visų atpažintų objektų kiekio santykis.
- F1-įvertis, apskaičiuojamas pagal formulę:

$$F1 = 2 * \frac{precision * recall}{precision + recall} \quad (3)$$

Kuo apskaičiuoto F1 įverčio rezultatas artimesnis vienetui, tuo aukštesnė gretinimo kokybė.

Primos eksperimento dalies metu tiriama kokios apimties tekstas bus gerai gretinamas naudojant statistinį sakinių sugretinimo komponentą. Buvo gauta, kad teksto gretinimo kokybė priklauso nuo teksto apimties. Kol tekstą paduodame nedideliais segmentais, sudarytais iki 100 sakinių klaidų skaičius ir dažnis yra nedidelis. Padidinus duodamų sakinių skaičių, gretinimo tikslumas smarkiai sumažėja.

Pateikiamas grafikas (žiūr. 4 pav.) parodo kaip gretinimo F1 metrika kinta nuo teksto apimties.



4 pav. Gretinimo kokybės priklausomybė nuo segmento apimties

Gretinimo kokybė smarkiai krinta didinant sakinių skaičių nuo 0 iki 500. Nuo 500 iki 3000 sakinių, gretinimo kokybė prastėja, bet pokytis nėra toks staigus. Ties 3000 sakinių riba, gretinimo

kokybės metrika pasiekia 0.3 ir tada beveik nustoja kisti. Padavus visą, nesuskaldytą tekstą, gretinimo kokybė prasčiausia – 0.27.

Antroje eksperimento dalyje matuojame kiek ir kaip semantinės analizės pritaikymas gali pagerinti gretinimo kokybę, gretinant pilną, į segmentus nedalinamą knygą.

Pasirinktos metrikos buvo paskaičiuotos visoms išbandytoms ontologijų kombinacijoms. Gauti rezultatai pateikiami lentelėje (žiūr. 3 lentelę).

3 lentelė. Eksperimento rezultatai

Naudojamos ontologijos	F1
Statistinis metodas	0.27
SUMO	0.40
WordNet	0.46
FrameNet	0.43
SUMO + WordNet	0.47
SUMO + FrameNet	0.45
WordNet + FrameNet	0.51
SUMO + FrameNet + WordNet	0.51

Gauti rezultatai rodo, kad geriausiai ilgo teksto gretinimas atliekamas gretinant naudojant arba WordNet ir FrameNet ontologijas, arba visas tris – WordNet, FrameNet ir SUMO. Tada sugretinimo kokybė geriausia – pasiekama 0.51 F1 metrikos vertė. Prasčiausias rezultatas gaunamas nenaudojant ontologijų ir taikant statistinį metodą, taip gretinant tekstą sugretinimo kokybė prasta, F1 metrika tik 0.27.

Semantinę analizę atliekant tik su viena ontologija, geriausi rezultatai gaunami naudojant WordNet – 0.46 F1 vertė, prasčiausias naudojant tik SUMO – 0.4. Naudojant FrameNet rezultatai tik nedaug geresni už SUMO – 0.43.

SUMO naudojimas duoda prasčiausius rezultatus, taip nutinka dėl klaidų pasitaikančių šioje ontologijoje. Jei jau naudojame WordNet, SUMO neduoda jokios papildomos informacijos.

Rezultatai

Šio darbo metu buvo suformuluotas ilgų tekstų gretinimo sistemos metodą ir jo pagrindu sukurti prototipą paremtą žodžių semantinė analize, bei surinkta informacija iš literatūros šaltinių. Buvo nustatyta ar ir kaip semantinė analizė gerina ilgų tekstų gretinimą.

Darbo rezultatai:

1. Buvo atlikta bendra tekstų gretinimo principų ir algoritmų analizė.
2. Įvardijamos problemos išskylančios gretinant ilgos apimties tekstus ir yra pasiūlomi problemų sprendimo būdai.
3. Nustatoma kaip gretinimo tikslumas priklauso nuo gretinimo apimties.
4. Pasiūlomas sakinių ir žodžių gretinimo metodas naudojantis ontologijas.
5. Nustatoma kaip gretinimo tikslumas priklauso nuo įvairių ontologijų ir jų kombinacijų naudojimo žodžių reikšmėms surasti.
6. Naudojant naujai suformuluotą metodą yra sugretinama Artūro Konano Doilio „Baskervilių šuo“ grožinės literatūros knyga ir jos vertimas. Yra paruošiama sugretintų knygos dalių lentelė.

Išvados

Atlikus literatūros šaltinių analizę ir eksperimentą buvo padarytos išvados:

1. Statistiniai gretinimo algoritmai veikia gerai su mažos apimties teksta ir dokumentais, tačiau yra nepritaikyti ilgos apimties tekstų, knygų, gretinimui. Padavus nepravalytą ir gretinimui neparuoštą knygą, jos gretinimas bus netikslus dėl ilgame tekste pasitaikančių negretinamų elementų: antraščių, išnašų, vertėjo komentarų, skyrių, puslapių numeracijos. Tradiciniai gretinimo algoritmai yra nepritaikyti knygoms gretinti ir norint pagerinti jų gretinimą reiktų pašalinti nereikalingus teksto elementus naudojant reguliarias išraiškas ir taikyti naujus gretinimo metodus – atlikti semantinę analizę žodžių arba sakinių porų radimui.
2. Statistiniai teksto gretinimo algoritmai efektyviausi kai gretinama iki 100 sakinių ilgio teksto segmentus. Gretinant tokius mažos apimties tekstus gretinimo kokybė geriausia: F1 metrika siekia 0.8 – 1. Segmentų apimtį padidinus – padidėja blogai sugretinamų sakinių skaičius. Gretinant pilną grožinės literatūros knygą gretinimo kokybė nukrinta iki 0.27.
3. Visų ontologijų naudojimas gretinimui atlikti pagerino teksto gretinimo kokybę. Pritaikius semantinę analizę, teksto gretinimo kokybės matmuo pagerėjo – nuo 0.27 iki 0.51. Naudojant skirtingas ontologijas pasiekti F1 metrikos rezultatai: SUMO – 0.4; FrameNet – 0.43; WordNet – 0.46; Geriausias rezultatas pasiekimas kombinuojant kelias ontologijas kartu, tada gretinimo kokybės metrika F1 yra 0.51.
4. SUMO ontologijos naudojimas semantinės analizės sugretinimo uždaviniuose nepagerina sugretinimo tikslumo jei jau naudojame WordNet ontologiją. Tai atsitinka todėl, kad SUMO naudoja WordNet ryšiams su natūralios kalbos žodžiais surasti, tačiau jina nepatiksline žodžių sugretinimo įvesdama naujas ontologijos sąvokas.
5. Naudojant semantinę gretinimą galima įvertinti ontologijų kokybę. Jei ontologija sugebėtų 100% tiksliai, kaip ir žmogus, sugretinti knygas tai parodytų jų ryšių ir sąvokų pilnumą. Gauti semantinio gretinimo rezultatai priklauso nuo naudojamos ontologijos kokybės.
6. Pasiūlytas semantinio sugretinimo metodas tinka anglų kalbos parafrazių generavimui. Semantiškai sugretinus didelius kiekius grožinės literatūros kūrinių galima gauti geros kokybės parafrazių ir tekstyną.

ȘALTINIAI

- [BFL98] BAKER, Collin F.; FILLMORE, Charles J.; LOWE, John B. The berkeley framenet project. In: *Proceedings of the 17th international conference on Computational linguistics-Volume 1*. Association for Computational Linguistics, 1998. p. 86-90.
- [CA16] COZMA, Pantin; ANCA, Raluca. LANGUAGE AND MACHINES: THE HUMAN FACTOR WHERE TO?. *Studii de Știință și Cultură*, 2016, 12.3.
- [CHE10] CHEN, Desai, et al. SEMAFOR: Frame argument resolution with log-linear models. In: *Proceedings of the 5th international workshop on semantic evaluation*. Association for Computational Linguistics, 2010. p. 264-267.
- [Chr15] CHRIST, Oliver. *Computer-assisted natural language translation*. U.S. Patent No 8,935,148, 2015.
- [DQQ11] DING, Huafu; QUAN, Lili; QI, Haoliang. The Chinese-English bilingual sentence alignment based on length. In: *2011 International Conference on Asian Language Processing*. IEEE, 2011. p. 201-204.
- [DSC+10] DAS, Dipanjan, et al. SEMAFOR 1.0: A probabilistic frame-semantic parser. *Language Technologies Institute, School of Computer Science, Carnegie Mellon University*, 2010.
- [Fel05] FELLBAUM, Christiane. WordNet and wordnets. 2005.
[žiūrėta 2020-03-15]. Prieiga per internetą:
< <https://wordnet.princeton.edu/citing-wordnet> >
- [Fel12] FELLBAUM, Christiane. WordNet. *The encyclopedia of applied linguistics*, 2012.
- [FJP03] FILLMORE, Charles J.; JOHNSON, Christopher R.; PETRUCK, Miriam RL. Background to framenet. *International journal of lexicography*, 2003, 16.3: 235-250.
- [GC93] GALE, William A.; CHURCH, Kenneth W. A program for aligning sentences in bilingual corpora. *Computational linguistics*, 1993, 19.1: 75-102.
- [HY97] HARUNO, Masahiko; YAMAZAKI, Takefumi. High-performance bilingual text alignment using statistical and dictionary information. *Natural Language Engineering*, 1997, 3.1: 1-14.
- [YWZ10] YU, Xin; WU, Jian; ZHAO, Weina. Dictionary-based Chinese-Tibetan sentence alignment. In: *2010 International Conference on Intelligent Computing and Integrated Systems*. IEEE, 2010. p. 489-493.
- [KR93] KAY, Martin; RÖSCHEISEN, Martin. Text-Translation Alignment', *Computational Linguistics* 19, 121-142. *Google Scholar Google Scholar Digital Library Digital Library*, 1993.

- [Lau18] LAUKAITYTĖ, Ieva. Ilgų tekstų gretinimo algoritmai. 2018.
- [MSB+14] MANNING, Christopher D., et al. The Stanford CoreNLP natural language processing toolkit. In: *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*. 2014. p. 55-60.
- [NP03a] NILES, Ian; PEASE, Adam. Mapping WordNet to the SUMO ontology. In: *Proceedings of the IEEE International Knowledge Engineering Conference*. 2003. p. 23-26.
- [NP03b] NILES, Ian; PEASE, Adam. Linking Lixicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology. In: *Ike*. 2003. p. 412-416.
- [RKD06] RIMKUTĖ, Erika; KOVALEVSKAITĖ, Jolanta; DAUDARAVIČIUS, Vidas. Daugiakalbių tekstynų naudojimas ir taikymas. *Darbai ir dienos. Kaunas: Vytauto Didžiojo universitetas, 45 (2006)*, 2006.
- [STR04] STRAPPARAVA, Carlo, et al. Wordnet affect: an affective extension of wordnet. In: *Lrec*. 2004. p. 40.
- [SPA07] STEPHAN, Grimm€ st; PASCAL, Hitzler€ st; ANDREAS, Abecker€ st. Knowledge representation and ontologies. *Semantic Web Services: Concepts, Technologies, and Applications*, 2007, 51-105.
- [Šev03] ŠEVCENKO, Michal. Online presentation of an upper ontology. *Proc. of Znalosti 2003*, 2003.
- [Šil12] ŠILINSKIS, Vaidas, et al. *Semantinė paieška naudojant ribotus žodynus*. 2012. PhD Thesis. Vytauto Didžiojo universitetas.
- [TWC17] TREGASKIS, Richard; WELLS, Scott; CLEGG, Caden. *Apparatus, system, and method for computer aided translation*. U.S. Patent No 9,767,095, 2017.
- [Ver00] VÉRONIS, Jean. From the Rosetta stone to the information society. In: *Parallel text processing*. Springer, Dordrecht, 2000. p. 1-24.
- [WP14] WELSH, Susan; PRIOR, Marc. OmegaT for CAT beginners. *Recuperat el*, 2014, 27.
- [XMY15] XU, Yong; MAX, Aurélien; YVON, François. Sentence alignment for literary texts. *LiLT (Linguistic Issues in Language Technology)*, 2015, 12.
- [ZAM14] ZOTTI, Patrizia; MATSUMOTO, Riccardo Apolloni Yuji. Sentence Alignment of a Japanese-Italian Parallel Corpus. Towards a web-based Interface. 2014. [žiūrėta 2020-03-05]. Prieiga per internetą: <http://www.anlp.jp/proceedings/annual_meeting/2014/pdf_dir/P1-6.pdf>

Sąvokų apibrėžimai

Ontologija – aprašytos konkrečios srities (angl. domain) sąvokos bei jų ryšiai.

Leksikonas – dydis, apibūdinantis natūralios kalbos morfemas, žodžius kartu ir jų būdingus požymius.

Parafrazė - frazės, posakio arba didesnio teksto prasmės nusakymas kitais žodžiais.

Tekstynas (angl. corpus) – tam tikru būdu organizuota tekstų sanakaupa, pritaikyta kiekybinei ir (arba) kokybinei analizei, paprastai saugoma elektroniniu formatu.

Vertimų atmintis (angl. translation memory) - duomenų bazė sudaryta iš žodžių ir frazių, bei jų vertimų.

Semantinis žodynas – duomenų struktūra nusakanti ryšius tarp natūralios kalbos žodžių ir ontologijų sąvokų.

1 priedas. Realizuoto prototipo vartotojo sąsajos ekranas

Realizuoto prototipo vartotojo sąsajos ekranas (teksto fragmentas iš knygos „Baskervilių šuo“). Ekrane pateikti originalus angliškas teksto fragmentas (3 stulpelis), vertimas iš lietuvių į anglų (7 stulpelis) ir lietuvių kalbos teksto fragmentai (9 stulpelis). 4 ir 5 stulpeliai rodo kokios frazės ir žodžiai gali būti sugretinti.

7/20

SEM EN

SEM LT

EN

LT

MOVE TO EN1-EN2

Open SentForm

Some people without possessing genius have a remarkable people , without being brilliant themselves , have a wonderful

power power

of stimulating it . I confess , my dear fellow , that I am very much in your debt . He had never said as m

to emulate gentleness to others . Thank you , dear friend , I am extremely grateful to you . He has neve

1	1	Some	2	Some	Some		Some	1	Kai	DT	DT	some	some
2	2	people	2	people	people		people	2	kurie	NNS	NNS	people	people
3	4	without	2	without	without			3	žmonės,	IN	,	without	,
4	5	possessing	2	possessing	being		without	4		VBG	IN	possess	without
5	6	genius	2	genius	brilliant		being	5	patys	NN	VBG	genius	be
6	9	have	2	have	have		brilliant	6	nebūdami	VBP	JJ	have	brilliant
7	0	a	2				themselves	7	genialūs,	DT	PRP	a	themselve
8	11	remarkable	2	remarkable	wonderful		,	8		JJ	,	remarkable	,
9	12	power	2	power	power		have	9	turi	NN	VBP	power	have
10	0	of	2				a	10	nuostabią	IN	DT	of	a
11	0	stimulating	2				wonderful	11	galią	VBG	JJ	stimulate	wonderful
12	0	it	2				power	12	įžiebt	PRP	NN	it	power
13	18		2				to	13	genialių		TO		to
14	19		2		Thank		emulate	14	minčių	PRP	VB	I	emulate
15	0	confess	2				gentleness	15	kitiems.	VBP	NN	confess	gentleness
16	0		2				to	16	Prisipažįstu,		TO		to
17	0	my	2				others	17	brangusis	PRP\$	NNS	my	other
18	22	dear	2	dear	dear		,	18	drauge,	RB		dear	,
19	23	fellow	2	fellow	friend		Thank	19	aš	JJ	VB	fellow	thank
20	24		2		,		you	20	jums		PRP		you
21	0	that	2				,	21	esu	IN		that	,
22	25		2		I		dear	22	be	PRP	RB	I	dear
23	26	am	2	am	am		friend	23	galo	VBP	NN	be	friend
24	27	very	2	very	extremely		,	24	dėkingas.	RB		very	,
25	0	much	2				I	25	Jis	RB	PRP	much	I
26	28	in	2	in	grateful		am	26	dar	IN	VBP	in	be
27	0	your	2				extremely	27	niekados	PRP\$	RB	you	extremely
28	0	debt	2				grateful	28	taip	NN	JJ	debt	grateful
29	31		2				to	29	apie		TO		to
30	32	He	2	He	He		you	30	mane	PRP	PRP	he	you
31	0	had	2				,	31	nebuvo	VBD		have	
32	34	never	2	never	never		He	32	kalbėjęs,	RB	PRP	never	he
33	35	said	2	said	talked		has	33		VBN	VBZ	say	have
34	38	as	2	as	so		never	34	todėl	RB	RB	as	never
35	39	much	2	much	much		talked	35	nesiginiau,	JJ	VBN	much	talk
36	0	before	2				about	36		IN	IN	before	about
37	40		2				me	37	jaučiausi		PRP		I
38	0	and	2				so	38	pamalonintas	CC	RB	and	so
39	42		2		I		much	39		PRP	RB	I	much
40	0	must	2				,	40	juk	MD		must	,
41	0	admit	2				so	41	iki	VB	IN	admit	so
42	0	that	2				I	42	tol	IN	PRP	that	I
43	0	his	2				do	43	jis	PRP\$	VBP	he	do

Incomprehensible

Semantic_alignments_phrases

OpenSemTuples

OpenFrameNetTuples

INFERENCE_Tuples

9 9

12 12

Axioms

ChosePhrase

2 priedas. Žodžių semantinis gretinimas

Pirmame priede pateikto teksto fragmento žodžių fellow ir friend (lt. draugas) WordNet ir SUMO ontologijų semantinių klasių sąrašo pavyzdys. Kuo daugiau semantinių klasių sutampa, tuo labiau tikėtina, kad žodžiai yra gretinami.

	wordid	word_en	synset_num_orig	stan_gram	wordnet_word	synset_num	SumoLabel	sem_dist
1	19	fellow	09871364	JJ	fellow	09871364	SocialRole	0
2	19	fellow	09871364	JJ	lover	09622302	Human	1
3	19	fellow	09871364	JJ	person	00007846	Human	2
4	19	fellow	09871364	JJ	individual	00007846	Human	2
5	19	fellow	09871364	JJ	someone	00007846	Human	2
6	19	fellow	09871364	JJ	soul	00007846	Human	2
7	19	fellow	09871364	JJ	mortal	00007846	Human	2
8	19	fellow	09871364	JJ	somebody	00007846	Human	2
9	19	fellow	09871364	JJ	cause	00007347	Agent	3
10	19	fellow	09871364	JJ	causal_agency	00007347	Agent	3
11	19	fellow	09871364	JJ	causal_agent	00007347	Agent	3
12	19	fellow	09908025	JJ	fellow	09908025	Male	0
13	19	fellow	09908025	JJ	male	09624168	Male	1
14	19	fellow	09908025	JJ	male_person	09624168	Male	1
15	19	fellow	09908025	JJ	person	00007846	Human	2
16	19	fellow	09908025	JJ	individual	00007846	Human	2
17	19	fellow	09908025	JJ	someone	00007846	Human	2
18	19	fellow	09908025	JJ	soul	00007846	Human	2
19	19	fellow	09908025	JJ	mortal	00007846	Human	2
20	19	fellow	09908025	JJ	somebody	00007846	Human	2
21	19	fellow	09908025	JJ	cause	00007347	Agent	3
22	19	fellow	09908025	JJ	causal_agency	00007347	Agent	3
23	19	fellow	09908025	JJ	causal_agent	00007347	Agent	3
24	19	fellow	09935990	JJ	fellow	09935990	Human	0
25	19	fellow	09935990	JJ	associate	09816771	SocialRole	1
26	19	fellow	09935990	JJ	peer	09626238	SocialRole	2
27	19	fellow	09935990	JJ	equal	09626238	SocialRole	2
28	19	fellow	09935990	JJ	compeer	09626238	SocialRole	2
29	19	fellow	09935990	JJ	match	09626238	SocialRole	2
30	19	fellow	09935990	JJ	person	00007846	Human	3
31	19	fellow	09935990	JJ	individual	00007846	Human	3
32	19	fellow	09935990	JJ	someone	00007846	Human	3
33	19	fellow	09935990	JJ	soul	00007846	Human	3
34	19	fellow	09935990	JJ	mortal	00007846	Human	3
35	19	fellow	09935990	JJ	somebody	00007846	Human	3
36	19	fellow	09935990	JJ	cause	00007347	Agent	4

Results		Messages						
	wordid	word_en	synset_num_orig	stan_gram	wordnet_word	synset_num	SumoLabel	sem_dist
1	23	friend	09679708	NN	friend	09679708	Protestant	0
2	23	friend	09679708	NN	christian	09678009	Christian	1
3	23	friend	09679708	NN	religious_person	09628382	Deist	2
4	23	friend	09679708	NN	person	00007846	Human	3
5	23	friend	09679708	NN	individual	00007846	Human	3
6	23	friend	09679708	NN	someone	00007846	Human	3
7	23	friend	09679708	NN	soul	00007846	Human	3
8	23	friend	09679708	NN	mortal	00007846	Human	3
9	23	friend	09679708	NN	somebody	00007846	Human	3
10	23	friend	09679708	NN	cause	00007347	Agent	4
11	23	friend	09679708	NN	causal_agency	00007347	Agent	4
12	23	friend	09679708	NN	causal_agent	00007347	Agent	4
13	23	friend	09763784	NN	friend	09763784	SocialRole	0
14	23	friend	09763784	NN	person	00007846	Human	1
15	23	friend	09763784	NN	individual	00007846	Human	1
16	23	friend	09763784	NN	someone	00007846	Human	1
17	23	friend	09763784	NN	soul	00007846	Human	1
18	23	friend	09763784	NN	mortal	00007846	Human	1
19	23	friend	09763784	NN	somebody	00007846	Human	1
20	23	friend	09763784	NN	cause	00007347	Agent	2
21	23	friend	09763784	NN	causal_agency	00007347	Agent	2
22	23	friend	09763784	NN	causal_agent	00007347	Agent	2
23	23	friend	09785042	NN	friend	09785042	SocialRole	0
24	23	friend	09785042	NN	associate	09816771	SocialRole	1
25	23	friend	09785042	NN	peer	09626238	SocialRole	2
26	23	friend	09785042	NN	equal	09626238	SocialRole	2
27	23	friend	09785042	NN	compeer	09626238	SocialRole	2
28	23	friend	09785042	NN	match	09626238	SocialRole	2
29	23	friend	09785042	NN	person	00007846	Human	3
30	23	friend	09785042	NN	individual	00007846	Human	3
31	23	friend	09785042	NN	someone	00007846	Human	3
32	23	friend	09785042	NN	soul	00007846	Human	3
33	23	friend	09785042	NN	mortal	00007846	Human	3
34	23	friend	09785042	NN	somebody	00007846	Human	3
35	23	friend	09785042	NN	cause	00007347	Agent	4
36	23	friend	09785042	NN	causal_agency	00007347	Agent	4

3 priedas. Knygos gretinimas su OmegaT

Knygos „Baskervilių šuo“ teksto sugretinimo pavyzdys sakinių lygmenyje. Jame parodyta OmegaT v 4.0.1 sugretinimo pavyzdys. Matome, kad 19 sakinių tik viršutinis (pažymėtas žalios širdies ženklu) yra geras.

Keep	Source	Target
☒	"If you had told us of your own free will it would have been a different thing," said the baronet, "you only told us, or rather your wife only told us, when it was forced from you and you could not help yourself."	– Jei mums būtųėt pasisakę savo noru, būtų visai kas kita, – tarė baronetas.
☒	"I didn't think you would have taken advantage of it, Sir Henry--indeed I didn't."	– Bet jūs prisipažinote, teisingiau, jūsų žmona prisipažino tik tada, kai kito kelio jums nebebuvo.
☒	"The man is a public danger."	– Aš nemaniau, kad jūs tuo pasinaudosite, pone Henri, dievaži, nemaniau.
☒	There are lonely houses scattered over the moor, and he is a fellow who would stick at nothing.	– Jis pavojingas visuomenei.
☒	You only want to get a glimpse of his face to see that.	Čia, raiste, sodybos toli viena nuo kitos, o toks žmogus su niekuo nesiskaitys.
☒	Look at Mr. Stapleton's house, for example, with no one but himself to defend it.	Tai matyti, pažvelgus į jo veidą.
☒	There's no safety for anyone until he is under lock and key."	Va kad ir misteris Steipltonas – argi jis apsigintų, būdamas visuose namuose vienui vienas?
☒	"He'll break into no house, sir."	Niekas neturės ramybės, kol tas žmogus vėl neatsidurs už grotų.
☒	I give you my solemn word upon that.	– Jis nesilauš nė į vienus namus, pone.
☒	But he will never trouble anyone in this country again.	Duodu jums garbės žodį.
☒	I assure you, Sir Henry, that in a very few days the necessary arrangements will have been made and he will be on his way to South America.	Ir šiame krašte jis daugiau nė vieno nebepalies.
☒	For God's sake, sir, I beg of you not to let the police know that he is still on the moor.	Patikėkite manimi, pone Henri, kad po kelių dienų viskas bus sutvarkyta, ir jis išplauks į Pietų Ameriką.
☒	They have given up the chase there, and he can lie quiet until the ship is ready for him.	Dėl Dievo, pone, labai prašau, nepraneškite policijai, kad jis dar slapstosi.
☒	You can't tell on him without getting my wife and me into trouble.	Jie liovėsi jo čia ieškoję, ir jis gali ramiai palaukti garlaivio.
☒	I beg you, sir, to say nothing to the police."	Jei pranešit apie jį, įstumsit į bėdą ir mane su žmona.
☒	"What do you say, Watson?"	Labai prašau, pone, nepranešti policijai.
☒	I shrugged my shoulders.	– Ką pasakysit, Watsonai?
☒	"If he were safely out of the country it would relieve the tax-payer of a burden."	Aš gūžtelėjau pečiais.
☒	"But how about the chance of his holding someone up before he goes?"	– Jei jis išvyks iš mūsų krašto, mokesčių mokėtojams viena našta bus mažiau!
☒	"He would not do anything so mad, sir."	– Bet tai nėra išsivadavimas, pone, kaip man atrodo?

Step 1: Adjust alignment parameters

Comparison Mode: Average Score: 0.877

Algorithm: Calculator: Counter:

☒ Segment ☐ Remove Tags ☒ Highlight

4 priedas. Semantinis sakinių gretinimas

Sukurto prototipo rezultato pavyzdys: sakinių sugretinimas atliktas naudojant semantinius žodynus. Pirmas stulpelis – anglų kalbos teksto fragmentai; antras stulpelis – atitinkami lietuvių kalbos teksto fragmentai; trečias stulpelis - teksto fragmentai gauti naudojant automatinę vertimo sistemą Google Translate ir verčiant lietuvių kalbos fragmentus į anglų kalbą (vertimas atliktas nuo 2020-03-10 iki 2020-03-15).

Results		Messages		
	id	EN_ORIG	LT	EN_TRANSLATE
1	7594	The Hound of the Baskervilles Chapter 1 Mr. Sherl...	BASKERVILIŲ ŠUO 1 skyrius PONAS ŠERLOKAS HOLMS...	BASKERVILLE SHOE Section 1 PONA SHERLOCK HOLMS ...
2	7595	Let me hear you reconstruct the man by an examin...	Nusakykite iš lazdos, kas jis per žmogus. Man atrodo, pradėj...	Say from a stick who he is through a man. It seems to me that I ...
3	7596	Some people without possessing genius have a re...	Kai kurie žmonės, patys nebūdami genialūs, turi nuostabią g...	Some people, without being brilliant themselves, have a wonde...
4	7597	Not that you are entirely wrong in this instance. The...	Tačiau šį kartą nelabai teapsinkote. Tas žmogus tikrai yra kai...	But this time, you're not too disappointed. That person is really ...
5	7598	We know there has been a presentation. We belie...	Kad tai dovana, mes žinome. Be to, spėjame, kad tarnyba ...	That's a gift, we know. In addition, we assume that the service ...
6	7599	From my small medical shelf I took down the Medic...	Iš nedidelės savo lentynos su medicinos knygomis išsiėmiau ...	From a small shelf with medical books I traveled to the medical ...
7	7600	And the dog? Has been in the habit of carrying this ...	O šuo? Šuo paprastai neša lazda paskui šeimininką. Kadangi...	O dog The dog usually wears a stick next to the owner. Since t...
8	7601	He was a very tall, thin man, with a long nose like ...	Jis buvo labai aukštas ir laibas, ilga nosimi; ta nosis nelygina...	He was very tall and had a long nose; his nose was like a pair ...
9	7602	It was necessary to make a home of my own. Come...	Reikėjo turėti savo namus. Matot, matot, mes ne taip smarki...	I needed to have my own home. Seeing, seeing, we are not so ...
10	7603	I observe from your forefinger that you make your o...	Iš jūsų smiliaus atrodo, kad pats sukātės cigaretės. Nesivar...	From your face it looks like you're turning cigarettes around. Fe...
11	7604	I trust, sir, that I have not inadvertently Just a little, ...	Tikiuosi, pone, kad nebūsiu netyčiomis. Šiek tiek, atsakė Hol...	I hope, sir, that I will not be inadvertent. A little bit, Holmes ans...
12	7605	I may say that I was his personal friend as well as hi...	Galiu pasakyti, kad serui Čarlzui aš buvau ne tik gydytojas, ...	I can say that for Sir Charles I was not only a doctor but also a f...
13	7606	Holmes leaned back in his chair, placed his finger-ti...	Holmsas atsilošė krėslė, sudūrė pirštų galus ir užsimerkė, lyg ...	Holmes leaned back on the chair, folded his fingers and closed...
14	7607	This, in truth, his neighbours might have pardoned...	Kaimynai, aišku, jam už tai būtų atleidę, nes mūsų krašte irgi ...	Neighbors, of course, would have forgave him for it, because in...
15	7608	It chanced that some little time later Hugo left his g...	O Hugas, ilgai nelaukęs, paliko svečius, norėdamas nunešt...	And Hugo, having not waited for long, left the guests to take th...
16	7609	But at length some sense came back to their craze...	Paskui, kai pakvaišę protai kiek prašviesėjo, visi tryliką, kie...	Then, when some crazy minds asked for it, all thirteen, as it wer...
17	7610	Riding slowly in this fashion they came at last upon ...	Taip pamažėle įdami, jie pagaliau pamatė šunis. Šie, nors g...	As they slowly started, they finally saw the dogs. These, though...
18	7611	Such is the tale, my sons, of the coming of the hou...	Tokia tad yra sakmė, vaikai mano, apie šunį, kuris nuo to lai...	This is the story of my children about a dog, which has since br...
19	7612	Now, Mr. Holmes, we will give you something a littl...	O dabar, pone Holmsai, paskaitysime kiek naujesnį atsitikim...	And now, Mr Holmes, let's recount a slightly more recent incide...
20	7613	It is only two years since he took up his residence a...	Baskervilių dvare jis teis gyveno dvejus metus, bet jau visi žin...	He lived for two years in the Basque Country, but everyone alre...
21	7614	Dr. James Mortimer, the friend and medical attend...	Tą patį patvirtino ir daktaras Džeimsas Mortimeris, velionio dr...	The same was confirmed by Dr. James Mortimer, the late friend ...