ԼԵԶՎԱԿԱՆ ՄՈԴԵԼՆԵՐԻ ԿԻՐԱՌՄԱՆ ՀԵՌԱՆԿԱՐՆԵՐԸ ՏԵՔՍՏԻ ԻՆՔՆԱՏԻՊՈՒԹՅԱՆ ԱՍՏԻՃԱՆԻ ԳՆԱՀԱՏՄԱՆ ԲԱՆԱԿԱՆ ՀԱՄԱԿԱՐԳԵՐՈՒՄ

https://doi.org/10.59982/18294359-25.1-gp-22

Գևորգ Արմենի Պետրոսյան
Ասպիրանտ
ՀԱՊՀ, ՏՀՏԷ ինստիտուտ
Տեղեկատվական տեխնոլոգիաների և ավտոմատացման ամբիոն
gapetrosyan14@gmail.com

Ամփոփագիր

Աշխատանքում ներկայացված է տեքստի ինքնատիպության աստիճանի որոշման բանական համակարգերում լեզվական մոդելների կիրառման հնարավոր մոտեցումների ուսումնասիրություն: Տեքստի ինքնատիպության աստիճանը որոշելու համար նախ և առաջ անհրաժեշտ է բացահայտել դրանում պարունակվող բոլոր փոխառված մասերը (ներառյալ միջլեզվական): Միջլեզվական փոխառությունների հայտնաբերումը ենթադրում է տարբեր լեզուներով գրված տեքստերի իմաստի համադրում, քանի որ ուղիղ թարգմանությունը սովորաբար չի արտահայտում տեքստի լեզվական առանձնահատկությունները: Աշխատանքում նշված լեզվական մոդելների կիրառման հնարավոր փոփոխություններն ու հարմարեցումները հետագայում կօգտագործվեն միջլեզվական փոխառությունների բացահայտման և տեքստի ինքնատիպության աստիճանի որոշման առաջարկվող երկփուլ համակարգի նախագծման և մշակման համար: Միջլեզվային փոխառությունների հայտնաբերումը հրատապ խնդիր է համացանցում հատկապես քիչ ներկայացված լեզուների համար, ինչպիսին է՝ հայերենը: Միջլեզվական փոխառությունների հայտնաբերման գործընթացում հիմնականում կիրառվում են բնական լեզվի մշակման վրա հիմնված մեթոդները, քանի որ այդ մեթոդները թույլ են տալիս ավելի խորը մակարդակով վերլուծել բնական լեզվով գրված տեքստերը: Աշխատանքում ներկայացված է նաև լեզվական մոդելների կիրառման միջոցով տեքստի ինքնատիպության աստիճանի գնահատման և նրանում պարունակվող փոխառությունների (ներառյալ միջլեզվական) բացահայտման առաջարկվող համակարգի աշխատանքի գծապատկերը:

Հիմնաբառեր: *Բնական լեզվի մշակում, միջլեզվական փոխառություն, լեմմատիզացիա, գրագողություն, տեքստի ինքնատիպություն, լեզվական մոդել:*

Ներածություն

Տեքստերի նմանության չափումը բնական լեզվի մշակման առաջադրանքների հիմքն է: Այն կարևոր դեր է խաղում տեղեկատվության որոնման, խմբավորման, մեքենայական թարգմանության, երկխոսության համակարգերի և փաստաթղթերի համապատասխանության որոշման մեջ [Wang and Dong]:

Համացանցային տեխնոլոգիաների զարգացմանը զուգընթաց օտար լեզուներով գրված տեքստերը դարձել են ավելի մատչելի, քան երբևէ: Բացի այդ, մեքենայական թարգմանության համակարգերի զարգացումը զգալիորեն մեծացրել է տեքստերը տարբեր լեզուներով արագ և ճշգրիտ թարգմանելու հնարավորությունը: Այս ամենը նպաստել է միջլեզվական տեքստային փոխառությունների տարածմանը:

Կատարվել են գիտահետազոտական աշխատանքներ՝ միալեզու տեքստերում առկա փոխառությունների հայտնաբերման խնդրի շուրջ [Саакян и др.]:

Ի տարբերություն միալեզու փոխառությունների, միջլեզվական փոխառությունների ժամանակ կասկածելի տեքստը և սկզբնաղբյուրը գրված են տարբեր լեզուներով, այսինքն՝ բնօրինակ տեքստը փոխառվում է թարգմանության միջոցով:

Միջլեզվային փոխառությունների բացահայտումը կարևոր խնդիր է, իսկ դրա գլխավոր ուղղություններից են՝ ակադեմիական ազնվության պահպանումը և մտավոր սեփականության պաշտպանությունը:

Գոյություն ունեն թարգմանության միջոցով տեքստը սկզբնաղբյուր լեզվից թիրախային լեզվի վերածելու տարբեր եղանակներ:

Այլ կերպ ասած՝ լեզուների միջև փոխառությունները կարող են դրսևորվել տարբեր ձևերով՝ պարզ թարգմանություն, թարգմանություն և վերաձևակերպում, նախադասությունների միաձուլում, թիրախային լեզվով գրված նախադասության բաժանում երկու կամ ավելի նախադասությունների և միաձուլում թարգմանությունից հետո [Mohtaj and Asghari]:

Նախադասությունների վերաձևակերպումը և թարգմանությունը կարելի է դիտարկել որպես կապակցված խնդիրներ՝ բնական լեզվի մշակման ոլորտում: Թարգմանությունը իմաստի պահպանումն է, երբ այն փոխանցվում է մեկ այլ լեզվի բառերով, իսկ վերաձևակերպումը իմաստի պահպանումն է՝ օգտագործելով միևնույն լեզվի այլ բառեր [Callison-Burch]:

Եթե տեքստը ենթարկվել է զգալի իմաստային և շարահյուսական փոփոխությունների, տեքստերի համեմատության վիճակագրական մեթոդները չեն ապահովում բավարար արդյունքներ: Նման փոփոխությունները բացահայտելու համար անհրաժեշտ են լեզվական մեթոդներ, որոնք ունակ են իրականացնել տեքստի ավելի խորը վերլուծություն [Chong]:

Հիմնախնդրի նկարագրությունը

Բնական լեզվի մշակումը (Natural Language Processing - NLP) արհեստական բանականության և մաթեմատիկական լեզվաբանության ոլորտում

հետազոտությունների ուղղություն է, որն ուսումնասիրում է համակարգչի միջոցով բնական լեզվի (խոսքի և տեքստերի) ընկալման, վերլուծության և սինթեզի խնդիրները [Arumugam and Shanmugamani]:

Համացանցում բազմալեզու բովանդակության ընդլայնումը նոր մարտահրավերներ է ստեղծում բնական լեզվի մշակման համակարգերի համար, քանի որ նրանք պետք է կարողանան մշակել տարբեր լեզուների շարահյուսական, ձևաբանական, իմաստաբանական և այլ առանձնահատկությունները: Այս մարտահրավերների հաղթահարումը պահանջում է լեզվաբանների և մեքենայական ուսուցման փորձագետների միջև միջառարկայական համագործակցություն՝ ստեղծելու բնական լեզվի մշակման մեթոդներ, որոնք կարող են արդյունավետ կերպով հաղթահարել բազմալեզու համատեքստում առաջացող բարդությունները:

Լեզվական մոդելները բնական լեզվի մշակման ամենահեռանկարային գործիքներից են: Լեզվական մոդելը վիճակագրական մոդել է, որը նախատեսված է տեքստում բառերի հաջորդականության հանդիպման հավանականությունը կանխատեսելու համար:

Լեզվական մոդելը մոդելավորում է բառերի $P(\mathrm{w}_1, \mathrm{w}_2, \ldots, \mathrm{w}_N)$ հաջորդականության հավանականության բաշխումը, այսինքն՝ թույլ է տալիս ասել, թե ինչ հավանականությամբ կարող է հանդիպել բառերի $\mathrm{w}_1, \mathrm{w}_2, \ldots, \mathrm{w}_N$ հաջորդականությունը: Լեզվական մոդելավորում է կոչվում $P(\mathrm{w}_N \mid \mathrm{w}_1, \mathrm{w}_2, \ldots, \mathrm{w}_{N-1})$ հաջորդականությունում հաջորդ բառը կանխատեսելու խնդիրը [Куратов]:

$$P(\mathrm{w}_1, \mathrm{w}_2, \ldots, \mathrm{w}_N) = P(w_1)P(\mathrm{w}_2|\mathrm{w}_1). \ldots P(\mathrm{w}_N \mid \mathrm{w}_1, \mathrm{w}_2, \ldots, \mathrm{w}_{N-1})$$

Բնական լեզվի մշակման ոլորտում մեքենայական և խորը ուսուցման մեթոդները լայնորեն կիրառվում են տարբեր առաջադրանքներում, ինչպիսիք են մեքենայական թարգմանությունը, տեղեկատվության որոնումը, հատուկ անունների ճանաչումը (Named Entity Recognition - NER), տեքստի քերականության և ուղղագրության ավտոմատ ստուգումը և այլն: Այնուամենայնիվ, բավարար արդյունավետություն ցուցաբերելու համար դրանք սովորաբար պահանջում են մեծ ծավալի նախապես մշակված տվյալներ [Куратов]:

Լեզվական մոդելները հաջողությամբ օգտագործվում են կիրառական բազմաթիվ ոլորտներում, ինչպիսիք են՝ խոսքի ճանաչումը և բնական լեզվի վիճակագրական մշակումը:

Վերջին տարիներին ամենատարածված լեզվական մոդելները դարձել են նեյրոնային ցանցերի վրա հիմնված մոդելները: Դրանք սովորում են մեծ քանակությամբ տեքստային տվյալների վրա և կարող են մշակել ավելի բարդ լեզվական կառուցվածքներ, ինչպիսիք են՝ բարդ նախադասությունները, փոխաբերությունները և այլն:

Աշխատանքի նպատակն է տեքստի ինքնատիպության աստիճանի գնահատման համակարգերում լեզվական մոդելների կիրառման հնարավոր մոտեցումների

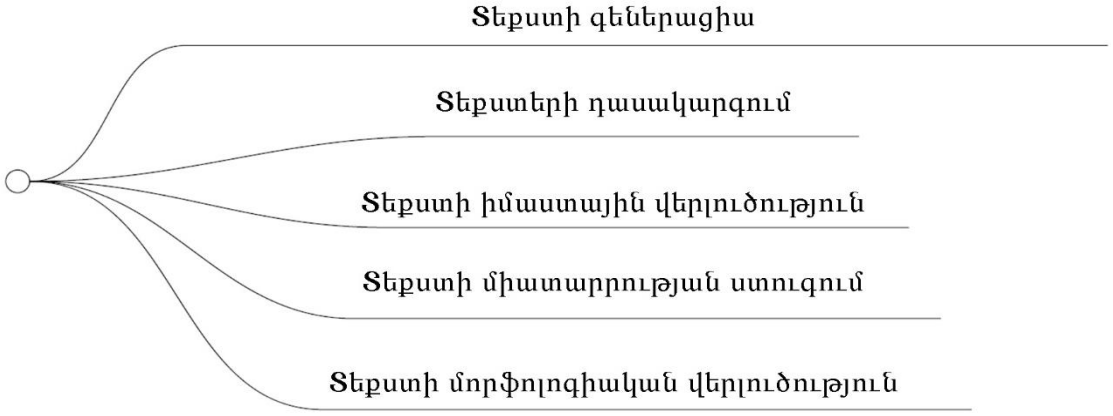
ուսումնասիրությունը և դրա հիման վրա տեքստի ինքնատիպության աստիճանի գնահատման առաջարկվող համակարգի աշխատանքի սխեմայի նախագծումը:

Տեքստի ինքնատիպության աստիճանի գնահատման բանական համակարգերում լեզվական մոդելների կիրառման հնարավորությունների ուսումնասիրություն

Ինքնատիպ ասելով հասկանում ենք տեքստի այն հատկանիշը, որը ենթադրում է, որ այն փոխառված չէ այլ հեղինակների աշխատանքներից կամ արտագրված չէ՝ ենթարկելով որոշակի փոփոխություններ: Այսինքն, ինքնատիպ է տեքստի այն մասը, որն ամբողջությամբ արտահայտում է հեղինակի սեփական մտքերը: Փոխառությունը իմաստային առումով ավելի լայն հասկացություն է: Տեքստի տվյալ մասը համարվում է փոխառված, եթե պարունակում է այլ աշխատանքներից հատվածներ՝ համապատասխան հղումներով այդ աշխատանքների վրա: Փոխառություն և ինքնատիպություն հասկացությունները իրենցից ներկայացնում են միևնույն երևույթի իրար փոխլրացնող կողմեր [Սահակյան, Պետրոսյան]:

Հետևաբար, տեքստի ինքնատիպության աստիճանը որոշելու համար առաջին հերթին անհրաժեշտ է բացահայտել դրանում պարունակվող բոլոր փոխառված մասերը: Լեզվական մոդելները տեքստի ինքնատիպության աստիճանի որոշման համակարգերում կարող են ունենալ բազմաթիվ կիրառություններ (Տե՛ս, նկար 1.)[1]:

Նկար 1.



Տեքստային փոխառությունների (ներառյալ միջլեզվական) բացահայտելու համար լեզվական մոդելների հնարավոր կիրառությունները

Տեքստի գեներացիա

Տեքստի գեներացիան, որը նաև հայտնի է որպես բնական լեզվի գեներացիա, բնական լեզվի մշակման կարևորագույն ոլորտներից է: Դրա նպատակն է տարբեր

[1] «Նկար 1»-ում ներկայացված են տեքստի ինքնատիպության աստիճանի գնահատման համակարգերում լեզվական մոդելների՝ հեղինակի կարծիքով առավել կարևոր կիրառությունները:

տեսակի մուտքային տվյալներից (տեքստ, պատկեր, աղյուսակ և գիտելիքների բազա և այլն) ստեղծել արժանահավատ և ընթեռնելի տեքստ՝ մարդկային լեզվով [Junyi et al.]:

Լեզվական մոդելները կարող են ֆիքսել համատեքստը, խոսքի ոճը և բառերի միջև փոխհարաբերությունները: Լեզվական մոդելների մեծ մասը, ինչպիսիք են՝ GPT մոդելները, սովորում են կանխատեսել հաջորդ բառի հայտնվելու հավանականության բաշխումը նախորդ բառերի հաջորդականության հիման վրա: Նման լեզվական մոդելները, որոնցում նեյրոնային ցանցի ընդունած որոշումները կախված են միայն ձախ կողմի բառերից, կոչվում են միակողմանի:

Այսպիսով, նախապես ուսուցանված մոդելները կարող են ստեղծել բնական լեզվին մոտ տեքստեր (տարբեր լեզուներով) և համեմատել դրանք տվյալների բազայում առկա տեքստերի հետ՝ հնարավոր փոխառությունները հայտնաբերելու նպատակով:

Տեքստերի դասակարգում

Տեքստի հետ կապված բազմաթիվ առաջադրանքներ հանգում են դասակարգման խնդրին՝ էմոցիոնալ երանգավորում, երկխոսության համակարգերում օգտագործողների մտադրությունների որոշում, վիրավորանքների և անպատշաճ խոսքի բացահայտում, տեքստի հեղինակների և թեմայի որոշում [Куратов]:

Ամբողջական տեքստի որոնման և տեքստերի դասակարգման առաջադրանքներում բառերի ամբողջական ձևաբանական վերլուծություն չի պահանջվում, այլ պահանջվում է միայն այն փաստի ստուգում, որ նշված երկու բառերը միևնույն բառի բառաձևեր են: Դա կարելի է անել լեմմատիզացիայի[2] միջոցով (բառերը բերելով սկզբնական տեսքին) կամ սթեմինգի[3] միջոցով, որը բաղկացած է բառերի որոշ անփոփոխ հատվածի առանձնացումից: Այնուամենայնիվ, մորֆոլոգիական վերլուծությունը, լեմմատիզացիան և սթեմինգը միշտ չէ, որ կարողանում են նույնականացնել իմաստով մոտ (մերձիմաստ) բառերը: Մերձիմաստ բառերի որոշման խնդիրը լուծվում է հատուկ թեզաուրուսների[4] միջոցով, որոնք կողմնորոշված գրաֆներ են, որոնցում գագաթները համապատասխանում են բառերին, իսկ կողերը՝ բառերի իմաստային կապերին: Երկու բառի նմանությունը որոշվում է գրաֆի երկու համապատասխան գագաթները միացնող ամենակարճ ուղու հիման վրա: Եթե անհրաժեշտ է հաշվի առնել բառերի համատեքստը, ապա խնդիրը շատ ավելի բարդանում է, և այն պետք է դասել տեքստի իմաստային մշակմանը [Белов и др.]:

Տեքստերի դասակարգումը կարող է օգտագործվել առաջարկվող երկփուլ համակարգի առաջին մակարդակում՝ տեքստերի մեծ հավաքածուից առավել փոքր ենթահավաքածու առանձնացնելու նպատակով, որոնցից ենթադրաբար փոխառված են դիտարկվող տեքստի որոշ հատվածներ, և որոնք կանցնեն ավելի մանրամասն

[2] (անգլ. lemmatization) բառի բազային կամ սկզբնական ձևը (լեմման) որոշելու գործընթաց
[3] (անգլ. stemming) բառի արմատը առանձնացնելու պարզեցված գործընթաց
[4] (անգլ. thesaurus) բառարան, որտեղ մերձիմաստ բառերը դասակարգված են ըստ խմբերի

վերլուծության՝ երկրորդ մակարդակում: Առաջին մակարդակում տեքստերի ֆիլտրումը զգալիորեն կնվազեցնի համակարգի աշխատանքի ժամանակը:

Տեքստի իմաստային վերլուծություն

Լեզվական մոդելները կարող են օգտագործվել տեքստի առանձին մասերի միջև իմաստային նմանությունները որոշելու համար: Նրանք կարող են ճիշտ մեկնաբանել բառերի իմաստները՝ կախված դրանց համատեքստից: Սա հատկապես օգտակար է համանունների դեպքում, երբ բառերը տարբեր համատեքստերում ունեն տարբեր իմաստներ:

Լեզվական մոդելները կարող են արտահայտել բառերի համատեքստից կախված իմաստը (լեզվից անկախ): Բառերի էմբեդինգները[5] ընդհանուր իմաստով կամայական չափանի վեկտորներ են: Նրանց օգնությամբ հնարավոր է չափել բառերի միջև իմաստային նմանությունը, քանի որ նման բառերի ներկայացումները բազմաչափ տարածությունում իրենցից ներկայացնում են բավական մոտ վեկտորներ: Նրանց միջոցով հնարավոր է նաև առանձնացնել միևնույն առարկային կամ երևույթին վերաբերող բառեր, քանի որ բազմաչափ վեկտորային տարածության մեջ նման բառերը կկազմեն խիտ դասավորված վեկտորների (կետերի) խումբ:

Տեքստի ոճական միատարրության սրուգում

Տեքստի ինքնատիպության աստիճանը որոշելու համար լեզվական մոդելները կարելի է կիրառել տեքստի վիճակագրական վերլուծության, նրանում առկա ոճական անհամապատասխանությունների կամ օգտագործված բառապաշարի հանկարծակի փոփոխության հայտնաբերման նպատակով: Տեքստի համար միատարրությունը նշանակում է ամբողջ տեքստի տրամաբանական հաջորդականության և միատեսակ ոճի պահպանում: Նախապես ուսուցանված մոդելները կարող են գնահատել տեքստերի միատարրությունը:

Եթե տեքստի որոշակի հատված ակնհայտորեն տարբերվում է տեքստի ընդհանուր ոճից, ապա մեծ է հավանականությունը, որ այդ հատվածը փոխառված է մեկ այլ աղբյուրից:

Տեքստի մորֆոլոգիական վերլուծություն

Բնական լեզվով տեքստի վերլուծության կարևոր փուլը նրա մորֆոլոգիական վերլուծությունն է, այսինքն՝ տեքստը կազմող լեքսեմաների[6] մասին մանրամասն մորֆոլոգիական տվյալներ ստանալը: Տեքստի մորֆոլոգիական վերլուծության և վիճակագրական մշակման ընթացակարգի ավտոմատացումը կիրառական լեզվաբանության արդիական խնդիրներից է [Минец, Горушкина]:

Տեքստի մորֆոլոգիական վերլուծությունը տեղեկատվություն է տրամադրում բառերի քերականական տարբեր հատկանիշների մասին (անկախ լեզվից), ինչպիսիք

[5] (անգլ. embedding) — օբյեկտի (բառի, արտահայտության, նախադասության և այլն) թվային ներկայացում վեկտորի տեսքով, որը արտացոլում է նրա իմաստային հատկանիշները

[6] (անգլ. lexeme) լեզվի հիմնական միավոր, որը ներկայացնում է բառարանային բառը իր ձևերի և իմաստների ամբողջությամբ

են՝ խոսքամասային պատկանելությունը, սեռը, հոլովը, սկզբնական ձևը և այլն: Լեզվական մոդելների միջոցով կարող է բարելավվել բառերի մասին հիմնական տեղեկատվության ստացման գործընթացը՝ ավելի ճշգրիտ մորֆոլոգիական վերլուծության շնորհիվ:

Լեմմատիզացիան՝ բառը իր սկզբնական ձևին բերելու գործընթացն է: Լեզվական մոդելները կարող են բառերը բերել իրենց՝ համատեքստից կախված սկզբնական ձևին: Լեմմատիզացիայի օգնությամբ հնարավոր է կարգավորել ընդհանուր բառարանը, որը հիմք է հանդիսանում տեքստերին համապատասխանող վեկտորների ստացման համար, ինչի արդյունքում կկրճատվեն վեկտորային տվյալների չափերը, ինչն էլ իր հերթին կօգնի բարելավել արդյունքները: Տեքստերը որպես վեկտորներ ներկայացնելուց հետո կարելի է չափել դրանց միջև եղած նմանությունը՝ օգտագործելով նմանության չափանիշներից մեկը, ինչպիսիք են՝ Էվկլիդյան հեռավորությունը, կոսինուսային նմանությունը և այլն:

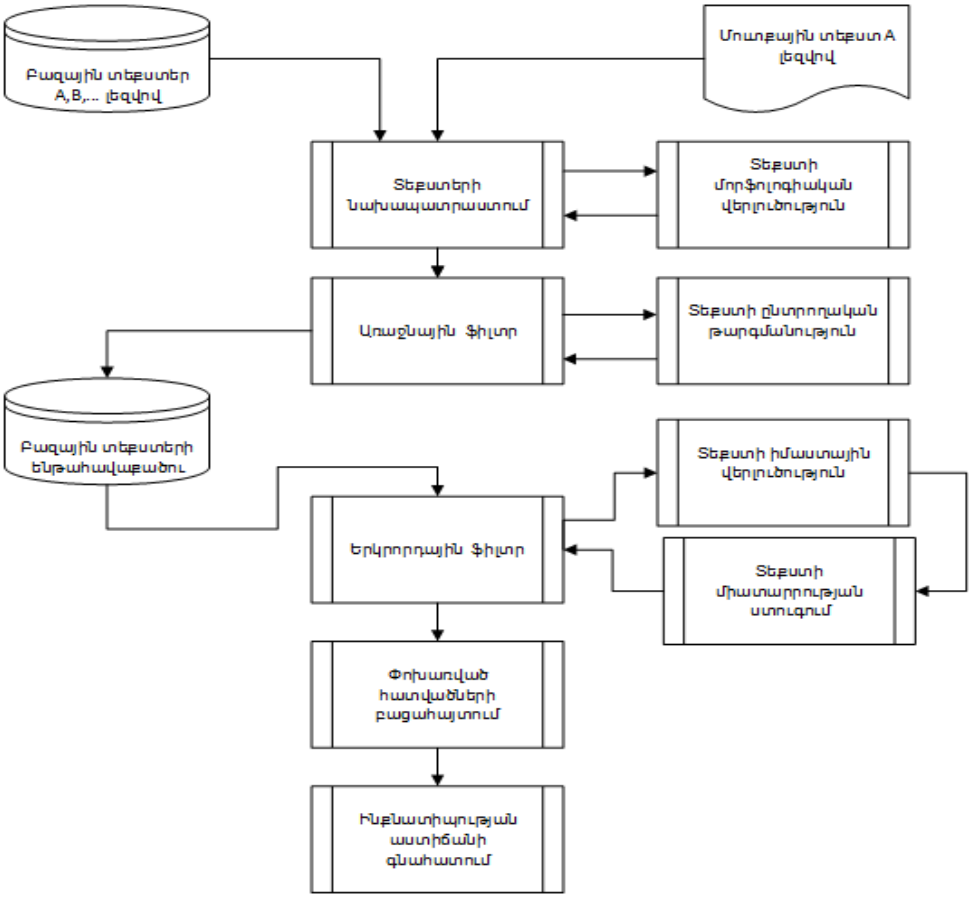
Տեքստի ինքնատիպության աստիճանի գնահատման բանական համակարգերում լեզվական մոդելների կիրառման հեռանկարները

Լեզվական մոդելների կիրառումը լայն հնարավորություններ է ստեղծում տեքստի ինքնատիպության գնահատման համակարգերի կատարելագործման համար: Դրանց օգտագործումը թույլ է տալիս բարձրացնել տեքստի վերլուծության ճշգրտությունը՝ բառերի և արտահայտությունների պարզ համեմատությունից անցնելով տեքստի իմաստաբանության և ոճի ըմբռնմանը: Դրանք օգնում են հայտնաբերել տեքստի վերաձևակերպումները, ճանաչել հոմանիշները և բառերի համատեքստը, ինչը փոխառությունների որոնումը դարձնում է ավելի ճշգրիտ: Հարկ է նաև նշել, որ լեզվական մոդելների հավելյալ ուսուցանումը (fine-tuning) որոշակի տեքստերի հավաքածուների վրա հնարավորություն է տալիս բարձրացնել արդյունավետությունը և ադապտացնել դրանց որևէ կոնկրետ ոլորտի տեքստերի համար, ինչպիսիք են գիտական, իրավական և այլ բնույթի տեքստերը: Լեզվական մոդելների կիրառման մեկ այլ կարևոր հեռանկարը համացանցում քիչ ներկայացված լեզուներով գրված տեքստերում փոխառությունների հայտնաբերման հնարավորությունն է՝ սահմանափակ տվյալների վրա սովորելու ունակության շնորհիվ:

Տեքստի ինքնատիպության աստիճանի գնահատման գործընթացի փուլային նկարագրությունը

Ստորև ներկայացված է տեքստի ինքնատիպության աստիճանի գնահատման և փոխառությունների (ներառյալ միջլեզվական) բացահայտման առաջարկվող գործընթացի փուլային նկարագրությունը՝ լեզվական մոդելների վերը նշված հնարավոր կիրառությունների միջոցով (Տե՛ս, նկար 2.):

Նկար 2.



Տեքստի ինքնատիպության աստիճանի գնահատման գործընթացի փուլային նկարագրությունը լեզվական մոդելների վերը նշված հնարավոր կիրառությունների միջոցով:

Ինչպես ցույց է տրված «Նկար 2»-ում, առաջնային ֆիլտրի կիրառումից առաջ տեքստերը անցնում են նախապատրաստման փուլ, ինչն իր մեջ ներառում է տեքստի մորֆոլոգիական վերլուծություն: Առաջնային ֆիլտրի արդյունքում բազային տեքստերից առաջանում է ավելի փոքր քանակով տեքստերի հավաքածու: Դրանք էլ անցնում են երկրորդային ֆիլտրի փուլ: Այս փուլում տեքստերը ենթարկվում են իմաստային վերլուծության և միատարրրության ստուգման, մուտքային տեքստում պարունակվող միջլեզվական փոխառությունները բացահայտելու համար: Տեքստի ընտրողական թարգմանության փուլը չի ենթադրում ամբողջական տեքստի թարգմանություն, այլ առանձին բառերի, արտահայտությունների կամ հատվածների թարգմանություն՝ առաջնային ֆիլտրի աշխատանքի ապահովման նպատակով:

Առաջարկվող համակարգի գործնական կիրառելիությունը հատկապես արդիական է բարձրագույն ուսումնական հաստատություններում, որտեղ անհրաժեշտ է հստակ գնահատել գիտահետազոտական աշխատանքների ինքնատիպությունը, կանխել

գրագողության դեպքերը և ապահովել կատարվող աշխատանքների պատշաճ ինքնատիպություն:

Եզրակացություն

Աշխատանքի շրջանակներում ներկայացված է լեզվական մոդելների կիրառման հնարավորությունը և հեռանկարները բազմալեզու միջավայրում տեքստի ինքնատիպության աստիճանի գնահատման խնդրում: Լեզվական մոդելների ճկուն կառուցվածքը թույլ է տալիս դրանք հարմարեցնել նույնիսկ համացանցում քիչ ներկայացված լեզուների համար՝ հաշվի առնելով այդ լեզուների կառուցվածքային առանձնահատկությունները: Աշխատանքի արդյունքում ձևավորվել են լեզվական մոդելների կիրառման առաջարկներ, որոնք հնարավոր կլինի ներդնել տեքստում առկա միջլեզվային փոխառությունների բացահայտման և տեքստի ինքնատիպության աստիճանի գնահատման համակարգերում: Աշխատանքում ներկայացված է նաև տեքստի ինքնատիպության աստիճանի գնահատման և փոխառությունների բացահայտման առաջարկվող համակարգի աշխատանքի գծապատկերը՝ լեզվական մոդելների վերը նշված հնարավոր կիրառությունների միջոցով:

Գրականության ցանկ

1. Սահակյան Ռ. Ռ, Պետրոսյան Գ. Ա. «Հետազոտական աշխատանքների ինքնատիպության աստիճանի գնահատման համակարգի նախագծում». Հայաստանի ճարտարագիտական ակադեմիայի լրաբեր, Հատոր 19, N1, 2022, էջ 98-103:
2. Wang J., Dong Y. «Measurement of Text Similarity: A Survey». Information, Volume 11, N9, 2020: https://doi.org/10.3390/info11090421.
3. Mohtaj S., Asghari G. «A Corpus for Evaluation of Cross Language Text Re-use Detection Systems». Journal of Information Systems and Telecommunication, Volume 10, N3, 2022, pp. (December 10, 2024).
   https://www.cis.upenn.edu/~ccb/publications/callison-burch-thesis.pdf.
4. Chong M. Y. M. «A Study on Plagiarism Detection and Plagiarism Direction Identification Using Natural Language Processing Techniques». University of Wolverhampton, 2013, 300 p. (January 10, 2025).
   https://wlv.openrepository.com/bitstream/handle/2436/298219/thesis.pdf?sequence=1
5. Arumugam R., Shanmugamani R. «Hands-On Natural Language Processing with Python». Packt Publishing, 2018. 312 p.
6. Junyi L., Tianyi T, Wayne X. Z, Jian-Yun N, Ji-Rong W. «Pre-trained Language Models for Text Generation: A Survey». Volume 1, N1, 2022, pp. 1-35, https://arxiv.org/pdf/2201.05273, (January 10, 2025).
7. Саакян Р. Р, Шпехт И. А, Петросян Г. А. «Нахождение наличия заимствований в научных работах на основе марковских цепей». Вестник Санкт-Петербургского университета. Прикладная математика. Информатика. Процессы управления, Том 19, Выпуск 1, 2023, с. 43-51: https://doi.org/10.21638/11701/spbu10.2023.104.

8. Куратов Ю. М. «Специализация языковых моделей для применения к задачам обработки естественного языка». Федеральное государственное автономное образовательное учреждение высшего образования «Московский физико-технический институт, 2020, 121 с.
9. Белов С. Д., Зрелова Д. П., Зрелов П. В., Кореньков В. В. «Обзор Методов Автоматической Обработки Текстов На естественном Языке». Сетевое научное издание «Системный анализ в науке и образовании», Выпуск 3, 2020, с. 8-22: https://doi.org/10.37005/2071-9612-2020-3-8-22
10. Минец Д. В, Горушкина А. В. «Морфологический анализ текста: Функциональные возможности». Litera, N3, 2017, с. 12-22: https://nbpublish.com/library_read_article.php?id=24112, (11 Декабря, 2024).

APPLICATION PERSPECTIVES OF LANGUAGE MODELS IN THE INTELLIGENT SYSTEMS FOR DETERMINING THE DEGREE OF TEXT UNIQUENESS

Gevorg Petrosyan
Post-Graduate Student
NPUA, Institute of ITTE, Chair of Information Technology and Automation
gapetrosyan14@gmail.com

Abstract

This paper provides an overview of possible approaches to using language models in intelligent systems for determining the degree of uniqueness of a text. To determine the degree of uniqueness of a text, it is necessary to first identify all borrowed parts contained in it (including cross-language ones). Finding cross-language borrowings involves comparing the meaning of the texts written in different languages, since a direct translation normally does not express the linguistic features of the text. Possible changes and adaptations of the language models applications mentioned in this work will be further used for the design and development of the proposed two-stage approach for identifying cross-language borrowings and determining the degree of uniqueness of the text. Detection of cross-language borrowings is an especially urgent task for languages that are underrepresented on the Internet, such as the Armenian language. Methods based on natural language processing are considered one of the highest priorities in the task of detecting cross-language borrowings, since they allow to analyze texts written in natural language at a deeper level. The work also presents the diagram of the proposed system for determining the degree of uniqueness of the text and identifying the borrowings (including cross-language ones) contained in it through the use of the mentioned applications of language models.

Keywords: *Natural language processing, cross-language borrowing, lemmatization, plagiarism, text uniqueness, language model.*

Ներկայացվել է՝ 03.04.2025թ.
Ուղարկվել է գրախոսման՝ 28.04.2025թ.