Հայոց Լեզվի Խոսքի Մասերի Պիտակավորում Թաքնված Մարկովյան Մոդելներով

Մանե Մխիթարյան, Ալբերտ Փիլիպոսյան, Էդուարդ Դանիելյան

15 դեկտեմբեր 2025

Նկարագրություն

Խոսքի մասերի պիտակավորումը բնական լեզվի մշակման գործընթաց է, որտեղ տեքստի յուրաքանչյուր բառին վերագրվում է համապատասխան խոսքի մաս: Այս խնդիրը հիմնականում լուծվում է լեզվական կանոններով, վիճակագրական մոդելներով և նեյրոնային ցանցերով: Այս ուսումնասիրության ընթացքում առաջարկել ենք Թաքնված Մարկովյան մոդելներով լուծում հայոց լեզվի համար, օգտագործելով Վիտերբիի և պոսթերիոր ալգորիթմները: Մեր առաջարկած մոտեցումը հասել է ավելի քան 92% ճշգրտության, որը լավագույնն է առկա հետազոտություններում: https://github.com/albpiliposyan/hidden_markov_model_pos_tagging

1 Ներածություն

Խոսքի մասերի պիտակավորումը կարևոր խնդիր է բնական լեզվի մշակման մեջ: Այն ենթադրում է տեքստի յուրաքանչյուր բառին ճիշտ քերականական խոսքի մաս (օրինակ՝ գոյական, բայ, ածական) վերագրելը՝ հիմնվելով դրա

իմաստի և կոնտեքստի վրա (Manning and Schütze, 1999; Kumawat and Jain., 2015): Այս խնդիրը կարևոր է, քանի որ արդյունքում ստացված պիտակները օգտագործվում են որպես հիմնական հատկանիշներ շատ այլ կիրառություններում, ինչպիսիք են շարահյուսական վերլուծությունը, մեքենայական թարգմանությունը և տեղեկատվության որոնումը: Խոսքի մասերի ճշգրիտ պիտակավորումը նաև օգնում է լուծել բառերի երկիմաստությունները, ինչն անհրաժեշտ է ավելի ընդլայնված լեզվական վերլուծության համար:

Խոսքի մասերի պիտակավորման մեթոդների մշակումն անցել է երեք հիմնական փուլով: Առաջին փուլում պիտակավորումը եղել է լեզվական կանոնների վրա հիմնված, օգտագործելով ձեռքով գրված լեզվական կանոններ և մեծ բառարաններ (Pham, 2020): Այս համակարգերը ճշգրիտ էին, բայց դժվարանում էին անծանոթ կամ երկիմաստ բառերի հետ աշխատել: Հաջորդ փուլը՝ վիճակագրական պիտակավորումը, լուծեց այդ խնդիրը՝ ներկայացնելով հավանականային մոդելներ, որոնք սովորել էին պիտակավորված տվյալների վրա: Հայտնի օրինակ է թաքնված Մարկովյան մոդելը, որը օգտագործում է և՛ բառի կոնկրետ խոսքի մաս լինելու հավանականությունը, և՛ մի խոսքի մասը մյուսին հաջորդելու հավանականությունը (Bărbulescu and Morariu, 2020): Վերջին տարիներին այս խնդրի լուծումը տեղափոխվել է խորը ուսուցման մոդելների դաշտ, ինչպիսիք են կրկնվող նեյրոնային ցանցերը և տրանսֆորմերների վրա հիմնված մոդելները, որոնք կարող են լեզվի ավելի բարդ օրինաչափություններ սովորել (Wang et al., 2015; Saidi et al., 2021):

Վիճակագրական թաքնված Մարկովյան մոդելի մեթոդը նշանակալի արդյունքների է հասել բազմաթիվ լավ ուսումնասիրված լեզուներում: Անգլերենի դեպքում, այս մոդելները, որոնք օգտագործում էին Վիտերբի ալգորիթմը ամենահավանական խոսքի մասերի հաջորդականությունը գտնելու համար, հասել էին ավելի քան 95% ճշգրտության այնպիսի տվյալների հավաքածուներում, ինչպիսին է Բրաունի կորպուսը (Aliwy et al., 2015): Նմանատիպ մոդելները լավ են աշխատել

նաև ձևաբանորեն հարուստ լեզուների համար՝ հասնելով մոտ 92% ճշգրտության: (Goyal et al., 2019). Չնայած ավելի նոր խորը ուսուցման մոդելները այժմ հասնում են նույնիսկ ավելի բարձր ճշգրտության՝ մինչև 97% (Akbik et al., 2018), թաքնված Մարկովյան մոդելը մնում է օգտակար և մեկնաբանելի բազային մոդել: Այն դեռևս արժեքավոր է հետազոտություններում՝ հատկապես ցածր ռեսուրսներով լեզուների համար, կամ որպես համեմատության մեկնարկային կետ:

Հայոց լեզվի համար խոսքի մասերի պիտակավորումը դժվար խնդիր է, քանի որ այն ունի սահմանափակ թվային ռեսուրսներ բնական լեզվի մշակման համար: Իրավիճակն ավելի է բարդանում խորը հետազոտությամբ պիտակավորված տվյալների բացակայության պատճառով, ի տարբերություն այնպիսի լեզուների, ինչպիսիք են անգլերենը կամ չինարենը: Չնայած այս դժվարություններին, ժամանակակից արևելահայերենի վերաբերյալ հետազոտությունները զարգանում են: Թաքնված Մարկովյան մոդելների նման վիճակագրական մոդելներ օգտագործող նախորդ ուսումնասիրությունները ցույց են տվել խոստումնալից արդյունքներ՝ հասնելով մոտ 81.25% ճշգրտության Վիտերբիի ալգորիթմի վրա հիմնված մեթոդով (Baghdasaryan, 2023):

Այս ուսումնասիրության նպատակն է վերագնահատել և վերլուծել ավանդական թաքնված Մարկովյան մոդելի աշխատանքը հայոց լեզվի համար՝ ստեղծելով ամուր և հուսալի հիմք ժամանակակից մոտեցումների հետ համեմատության համար:

2 Թաքնված Մարկովյան Մոդելներ

Թաքնված Մարկովյան մոդելը հավանականային մոդել է, որը ներկայացնում է համակարգ հիմնված Մարկովի պրոցեսի վրա: Այն ենթադրում է, որ համակարգի յուրաքանչյուր վիճակ կախված է միայն իր անմիջական նախորդող վիճակից, այլ ոչ թե ամբողջ նախորդող վիճակներից (Eddy, 2004): Այսինքն, եթե դիտարկենք վիճակի

փոփոխականների հաջորդականությունը $q_1, ...q_n$, ապա վերը նշված սահմանման մաթեմատիկական համարժեքը կլինի՝

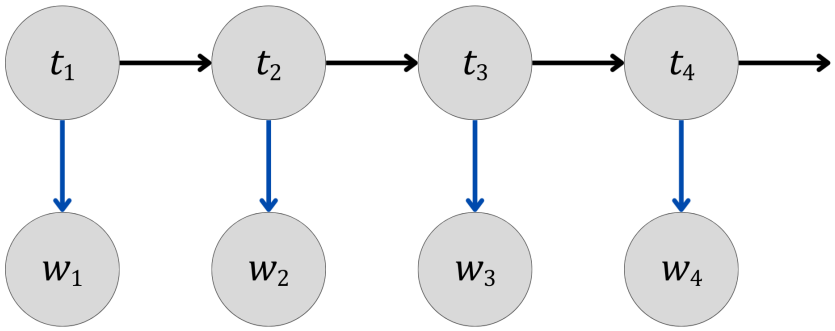
$$P(q_i = x|q_1...q_{i-1}) = P(q_i = x|q_{i-1}) \tag{1}$$

որը հայտնի է որպես առաջին կարգի Մարկովի շղթա:

Թաքնված Մարկովի մոդելը օգտակար է, երբ մեր հետաքրքրության առարկա հանդիսացող իրադարձությունները անմիջականորեն չեն դիտարկվում, այսինքն՝ դրանք թաքնված են: Հայտնի օրինակ է տեքստի խոսքի մասի պիտակավորման խնդիրը. այսինքն՝ տրված տեքստի դեպքում որոշել տվյալ տեքստի յուրաքանչյուր բառի խոսքի մասը: Այս խնդրի դեպքում խոսքի մասերը թաքնված են, դիտարկվում են միայն նախադասության բառերը:

Թաքնված Մարկովի մոդելը սահմանվում է հետևյալ բաղադրիչներով (Jurafsky and Martin, 2025):

- $Q = q_1 q_2 ... q_N$ - N վիճակների բազմություն,
- $A = a_{11} ... a_{ij} ... a_{NN}$ - A անցման հավանականության մատրից, որի յուրաքանչյուր a_{ij} էլեմենտը ներկայացնում է i վիճակից j վիճակին անցնելու հավանականությունը, այնպես որ $\sum_{i=1}^{N} a_{ij} = 1 \; \forall i$,
- $B = b_i(o_t)$ - դիտարկումների հավանականությունների հաջորդականություն (արտանետման հաջորդականություն), որտեղ յուրաքանչյուրն արտահայտում է այն հավանականությունը, որ o_t (վերցված $V = v_1, v_2, ..., v_V$ բառարանից) դիտարկումը ստեղծվել է q_i վիճակից,
- $\pi = \pi_1, ...\pi_N$ - վիճակների սկզբնական հավանականություններ: π_i-ն այն հավանականությունն է, որ Մարկովյան շարքը կսկսի i վիճակից: Որոշ



Նկար 1: Սև սլաքը ցույց է տալիս անցումային հավանականությունը, կապույտ սլաքը՝ արտանետման հավանականությունը:

վիճակներ j-երի դեպքում կարող ենք ունենալ $\pi_j = 0$, այսինքն՝ այդ վիճակները չեն կարող մեկնարկային լինել: Ինչպես նաև՝ $\sum_{i=1}^{n} \pi_i = 1$:

Առաջին կարգի թաքնված Մարկովյան մոդելի համար անհրաժեշտ է երկու ենթադրություն՝

1. Մարկովի ենթադրութոյուն:

$$P(q_i \mid q_1 ... q_{i-1}) = P(q_i \mid q_{i-1}) \tag{2}$$

2. Ելքի անկախություն:

$$P(o_i \mid q_1 ... q_T, o_1 ... o_T) = P(o_i \mid q_i) \tag{3}$$

որտեղ o_i-ը դիտարկում է՝ վերցված V բառարանից, իսկ T-ն մուտքային տվյալներում դիտարկումների ($O = o_1 o_2 ... o_T$) ընդհանուր թիվն է:

Խոսքի մասերի պիտակավորման խնդրում թաքնված Մարկովյան մոդելը ունի երկու բաղադրիչ՝ անցման և արտանետման հավանականություններ, որոնք գնահատվում են օգտագործելով ուսուցման համար նախատեսված պիտակավորված տվյալների կորպուս:

Անցման հավանականությունը ցույց է տալիս նախորդ խոսքի մասի առկայության դեպքում տվյալ խոսքի մասի առաջացման հավանականությունը: Պարզ ասած, եթե դիտարկվում են օժանդակ բայերը (օրինակ՝ է, եմ, էին և այլն), ավելի հավանական է, որ դրանք հաջորդեն կամ նախորդեն հիմնական բայերին (օրինակ՝ վազել, գնալ, գրել և այլն), ուստի ակնկալվում է, որ դրանց հավանականությունը ավելի բարձր կլինի: Մենք հաշվարկում ենք այս անցումային հավանականության առավելագույն ճշմարտանմանության գնահատիչը, այսինքն երբ մի խոսքի մասը հայտնվում է պիտակավորված կորպուսում, որքան է հավանականությունը որ այն հաջորդում է մյուս խոսքի մասին՝

$$P(t_i|t_{i-1}) = \frac{C(t_{i-1}, t_i)}{C(t_{i-1})} \tag{4}$$

Մյուս կողմից, $P(w_i, t_i)$ արտանետման հավանականությունը ցույց է տալիս w_i բառի t_i խոսքի մաս լինելու հավանականությունը: Արտանետման հավանականության առավելագույն ճշմարտանմանության գնահատիչը կլինի

$$P(w_i, t_i) = \frac{C(t_i, w_i)}{C(t_i)} \tag{5}$$

Այսպիսով՝ խոսքի մասերը թաքնված են, և մենք հետաքրքրված ենք դրանք գտնելով: Դրա համար մենք կատարում ենք վերծանում, որի ֆունկցիոնալությունն այն է, որ ընտրենք այն $t_1...t_n$ խոսքի մասերի հաջորդականությունը, որն ամենահավանականն է՝ հաշվի առնելով $w_1...w_n$ դիտարկումը:

$$\hat{t}_{1:n} = \underset{t_1,...,t_n}{\text{argmax}}\, P(t_1...t_n|w_1...w_n) \tag{6}$$

Նկատենք, որ վերը նշվածը պայմանական հավանականություն է, ուստի մենք կարող ենք օգտագործել Բայեսի կանոնը՝ $\hat{t}_{1:n}$-ի հաշվարկները հեշտացնելու համար: Այսպիսով, մենք ստանում ենք

$$\hat{t}_{1:n} = \underset{t_1,...,t_n}{\text{argmax}}\, \frac{P(w_1...w_n|t_1...t_n)P(t_1...t_n)}{P(w_1...w_n)} \tag{7}$$

Նկատենք, որ հայտարարը կախված չէ t-ից և, հետևաբար, կարող է դիտարկվել որպես հաստատուն: Քանի որ հաստատունով բազմապատկումը չի ազդում $argmax$-ի արդյունքի վրա, մենք կարող ենք հանել հայտարարը: Հավասարումը պարզեցվում է հետևյալ կերպ՝

$$\hat{t}_{1:n} = \underset{t_1,\ldots,t_n}{\text{argmax}}\, P(w_1 \ldots w_n | t_1 \ldots t_n) P(t_1 \ldots t_n) \tag{8}$$

Ավելին, օգտագործելով թաքնված Մարկովյան մոդելների երկու ենթադրությունները՝ 2-րդ և 3-րդ հավասարումները, խոսքի մասի պիտակավորման խնդրի համար կունենանք՝

$$P(w_1 \ldots w_n) \approx \prod_{i=1}^{n} P(w_i | t_i) \tag{9}$$

$$P(t_1 \ldots t_n) \approx \prod_{i=1}^{n} P(t_i | t_{i-1}) \tag{10}$$

Վերը նշված հավասարումները մոտավոր են, քանի որ նախադասություններում միանշանակ չէ բառի խոսքի մասի պիտակի կախվածությունը միայն իր անմիջական նախորդ խոսքի մասի պիտակից: Այս ենթադրությունները 6-րդ հավասարման մեջ տեղադրելով՝ ստանում ենք միայն նախորդ բառի խոսքի մասի պիտակից կախված ամենահավանական խոսքի մասի պիտակների հաջորդականության հավասարումը:

$$\hat{t}_{1:n} = \underset{t_1,\ldots,t_n}{\text{argmax}}\, P(t_1 \ldots t_n | w_1 \ldots w_n) \approx \underset{t_1,\ldots,t_n}{\text{argmax}} \prod_{i=1}^{n} P(w_i | t_i) P(t_i | t_{i-1}) \tag{11}$$

Այս խնդիրը լուծելու համար օգտագործվել է 3-րդ և 4-րդ ենթաբաժիններում նկարագրված ալգորիթմները:

3 Վիտերբիի ալգորիթմ

Այս ալգորիթմը սկզբնապես ստեղծվել էր փաթույթային կապի համակարգերի համար (Viterbi, 1967): Հետագայում ցույց տրվեց, որ ալգորիթմը կարող է

դիտարկվել որպես դինամիկ ծրագրավորման օրինակ, որի կառուցվածքը նման է նվազագույն խմբագրման հեռավորության ալգորիթմին (օրինակ՝ նվազագույնը քանի փոփոխությունով է հնարավոր մի բառից ստանալ մի ուրիշ բառ) (Omura, 1969): Ֆորնին նշել է, որ ալգորիթմն իսկապես առավելագույն ճշմարտանման է և, հետևաբար, միշտ օպտիմալ է այն իմաստով, որ այն միշտ գտնում է ամենաբարձր հավանականությամբ ուղին: (Forney, 1974)

Ալգորիթմը սկսում է հավանականության մատրից կազմելով: Մատրիցի յուրաքանչյուր սյուն համապատասխանում է մեկ o_t դիտարկման, և ամեն տող ներկայացնում է մեկ վիճակ, որում տվյալ դիտարկումը կարող է լինել: Այսպիսով, յուրաքանչյուր սյուն ունի հավանականություն ամեն q_i վիճակի համար միավորված մեկ թաքնված Մարկովյան մոդելի գրաֆում, որը սահմանված է որպես λ=(A, B, π)).

Մատրիցի յուրաքանչյուր էլեմենտ ցույց է տալիս հավանականությունը, որ առաջին t դիտարկումները տեսնելու և $q_1, ..., q_{t-1}$ վիճակների ամենահավանական հաջորդականությամբ անցնելու դեպքում, թաքնված Մարկովյան մոդելը կգտնվի j վիճակում: Յուրաքանչյուր $v_t(j)$-ի արժեք հաշվելու համար, ռեկուրսիվորեն ընտրվում է ամենավահանական ճանապարհը, որը կարող է հանգեցնել տվյալ էլեմենտին: Մաթեմատիկորեն՝

$$v_t(j) = \max_{q_1,...,q_{t-1}} P(q_1, ..., q_{t-1}, o_1, ...o_t, q_t = j \mid \lambda)$$

Հաշվի առնելով, որ $t - 1$ պահին յուրաքանչյուր վիճակում գտնվելու հավանականությունը արդեն հաշվված է, Վիտերբիի հավանականությունը հաշվվում է՝ վերցնելով այն ուղին, որն ամենամեծ հավանականությամբ կհանգեցնի տվյալ էլեմենտին: Այսպիսով, եթե դիտարկվում է t պահին q_t վիճակում գտնվելը, ապա $v_t(j)$ հաշվվում է

$$v_t(j) = \max_{i=1}^{N} v_{t-1(i)} a_{ij} b_j(o_t)$$

հավասարումով, որտեղ $v_{t-1}(i)$-ը Վիտերբիի նախորդ քայլում i-րդ վիճակին

համապատասխանող ուղու հավանականությունն է:

Բոլոր վիճակների և ժամանակների համար $v_t(j)$-ի արժեքը հաշվելուց հետո հնարավոր է որոշել ամենահավանական վիճակների հաջորդականության հավանականությունը: Վիճակների լավագույն հաջորդականությունը ստանալու համար պահվում է այն թաքնված վիճակների ուղիները, որոնք հանգեցնում են հաջորդ վիճակին, և վերջում հետագծվում է լավագույն ուղին մինչև սկիզբը (Վիտերբիի հետադարձ որոնում) հաշվողական $O(TN^2)$ բարդությամբ (Jurafsky and Martin, 2025): Լավագույն արժեքը սահմանվում է հետևյալ կերպ՝

$$P^* = \max_{i=1,\ldots,N} v_T(i)$$

4 Պոսթերիոր վերծանում

Մեկ այլ վերծանման ալգորիթմ թաքնված Մարկովյան մոդելներում պոսթերիոր վերծանումն է, որտեղ յուրաքանչյուր t դիրքի համար կանխատեսված պիտակը ամենաբարձր պոսթերիոր հավանականություն ունեցող վիճակն է, այլ ոչ թե ամենահավանական ընդհանուր հաջորդականությունը (ինչպես Վիտերբիի վերծանման դեպքում): Սա նվազագույն Բայեսի ռիսկի վերծանման հատուկ դեպք է Համինգի կորստի դեպքում (Gormley, 2023):

Այս վերծանումը շատ նման է ագահ պիտակավորմանը, քանի որ երկուսն էլ գտնում են պիտակի ամենաբարձր հավանականությունը յուրաքանչյուր t դիրքի համար: Այնուամենայնիվ, կա մի կարևոր տարբերություն՝ պոսթերիոր վերծանումը օգտագործում է գլոբալ տեղեկատվություն (այսինքն՝ այն աշխատում է ամբողջ նախադասության վրա), մինչդեռ ագահ պիտակավորումը հիմնված է միայն տեղային հավանականությունների վրա (Georgetown, 2016):

Այս գլոբալ որոշումները կայացնելու համար, պոսթերիոր վերծանումը

օգտագործում է առաջ շարժվող և հետադարձի ալգորիթմները (ենթաբաժին 4.1 և 4.2)՝ յուրաքանչյուր t դիրքի և j վիճակի համար, այդ վիճակում գտնվելու պոսթերիոր հավանականությունը հաշվարկելու համար՝ ունենալով ամբողջ O դիտարկման հաջորդականությունը՝

$$P(q_t = j \mid O) = \frac{\alpha_t(j)\beta_t(j)}{P(O \mid \lambda)}$$

որտեղ $\alpha_t(j)$-ն առաջ շարժվելու հավանականությունն է, $\beta_t(j)$-ն՝ հետադարձ հավանականությունը: Այսպիսով, T երկարությամբ հաջորդականության համար, որն ունի N թաքնված վիճակներ, ստացվում է $T \times N$ չափի մատրից, որտեղ յուրաքանչյուր տող համապատասխանում է t-րդ դիրքին, յուրաքանչյուր սյուն համապատասխանում է թաքնված q վիճակի, և մատրիցի յուրաքանչյուր տարր ներկայացնում է այն հավանականությունը, որ t դիրքում գտնվող բառը գտնվում է q վիճակում:

Վերջնական փուլում՝ պոսթերիոր վերծանումը ստանալու համար, ընտրվում է յուրաքանչյուր դիրքում ամենաբարձր հավանականություն ունեցող վիճակը:

$$\hat{q}_t = \underset{i}{\text{argmax}}\, P(q_t = j \mid O)$$

4.1 Առաջ շարժվող ալգորիթմ

Առաջ շարժվող ալգորիթմը հաշվարկում է հաջորդականության մինչև t դիրքը դիտարկելու և այդ դիրքում j վիճակում լինելու հավանականությունը:

$$\alpha_t(j) = P(o_1, o_2, ..., o_t, q_t = j) = \sum_{i=1}^{N} P(o_1, ...o_t, q_{t-1} = i, q_t = j) =$$

$$\sum_{i=1}^{N} P(o_t, q_t = j \mid o_1, ...o_{t-1}, q_{t-1} = i) P(o_1, ...o_{t-1}, q_{t-1} = i) =$$

$$\sum_{i=1}^{N} P(o_t \mid o_1, ...o_{t-1}, q_{t-1} = i, q_t = j) P(q_t = j \mid o_1, ...o_{t-1} q_{t-1} = i) P(o_1, ..., o_{t-1} q_{t-1} = i)$$

$$= \sum_{i}^{N} P(o_t \mid q_t = j) P(q_t = j \mid q_{t-1} = i) P(o_1, ..., o_{t-1}, q_{t-1} = i)$$

$$= \sum_{i=1}^{N} b_j(o_t) a_{ij} \alpha_{t-1}(i)$$

Ալգորիթմը պատկանում է դինամիկ ծրագրավորման ալգորիթմների կատեգորիային: Այն արդյունավետորեն հաշվարկում է j վիճակում ավարտվող բոլոր հնարավոր ուղիները՝ $O(TN^2)$ բարդությամբ (Gormley, 2023): Առաջ շարժվող ալգորիթմի ֆորմալ ռեկուրսիան հետևյալն է.

1. Սկզբնական կարգավորում

$$\alpha_1(j) = \pi_j b_j(o_1) \quad 1 \leq j \leq N$$

2. Ռեկուրսիա

$$\alpha_t(j) = \sum_{i=1}^{N} \alpha_{t-1}(i) a_{ij} b_j(o_t), \quad 1 \leq j \leq N, \quad 1 < t \leq T$$

3. Դադարեցում

$$P(O \mid \lambda) = \sum_{i=1}^{N} a_T(i)$$

Այս ալգորիթմը հաշվարկում է մինչ տվյալ պահը դիտարկումների հանդիպման հավանականությունը տվյալ դիրքում գտնվելու դեպքում:

4.2 Հետադարձ ալգորիթմ

Հետադարձ ալգորիթմը հաշվարկում է մնացած հաջորդականության՝ $t + 1$ դիրքից մինչև վերջ, դիտարկելու հավանականությունը ընթացիկ i վիճակում:

$$\beta_t(i) = P(o_{t+1}, o_{t+2}, ..., o_T \mid q_t = i, \lambda) = \sum_{j=1}^{N} P(q_{t+1} = j, o_{t+1}, ..., o_T \mid q_t = i) =$$

$$\sum_{j=1}^{N} P(o_{t+1} \mid q_{t+1} = j, q_t = i, o_{t+2}, ..., o_T) P(q_{t+1} = j \mid q_t = i, o_{t+2}, ..., o_T) P(o_{t+2}, ..., o_T \mid q_{t+1} = j) =$$

$$\sum_{j=1}^{N} P(o_{t+1} \mid q_{t+1} = j) P(q_{t+1} = j \mid q_t = i) P(o_{t+2}, ..., o_T \mid q_{t+1} = j) = \sum_{j=1}^{N} b_j(o_{t+1}) a_{ij} \beta_{t+1}(j)$$

1. Սկզբնական կարգավորում

$$\beta_T(i) = 1, \quad 1 \leq i \leq N$$

2. Ռեկուրսիա

$$\beta_t(i) = \sum_{i=1}^{N} a_{ij} b_j(o_{t+1}) \beta_{t+1}(j), \quad 1 \leq i \leq N, \quad 1 \leq t < T$$

3. Դադարեցում

$$P(O \mid \lambda) = \sum_{j=1}^{N} \pi_j b_j(o_1) \beta_1(j)$$

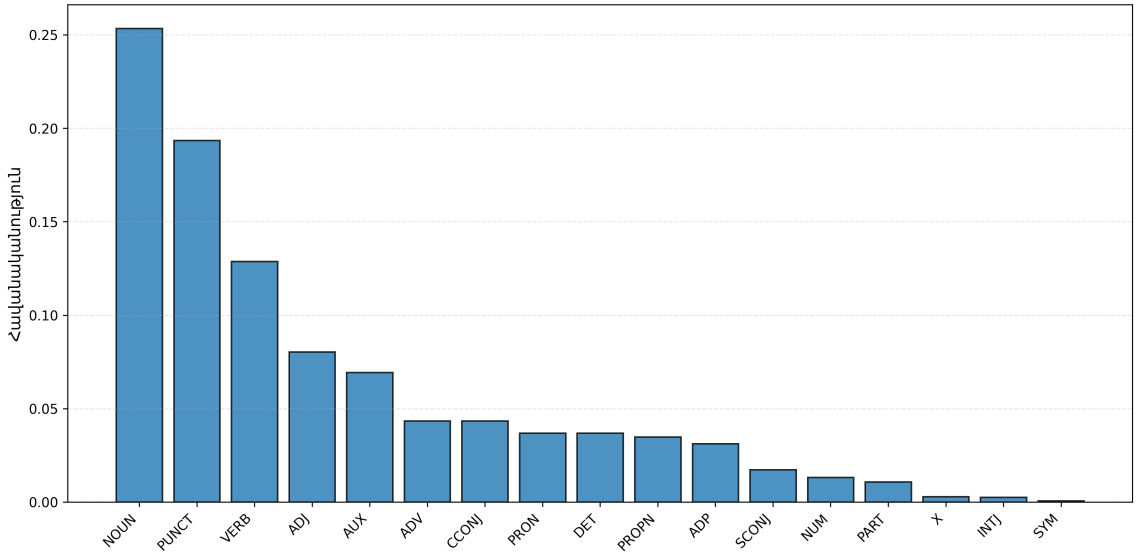
Այսպիսով, այս ալգորիթմը հաշվարկում է հաջորդող դիտարկումների հանդիպման հավանականությունը ընթացիկ դիրքում գտնվելիս:

5 Արդյունքներ

Այս հետազոտության համար օգտագործել ենք "Universal Dependencies treebank for Eastern Armenian" տվյալների կորպուսը (https://github.com/

UniversalDependencies/UD_Armenian-ArmTDP): Այն պարունակում է հայերեն նախադասություններ՝ մանրամասն լեզվական տեղեկատվությամբ, ներառյալ յուրաքանչյուր բառի համար խոսքի մասի պիտակը:

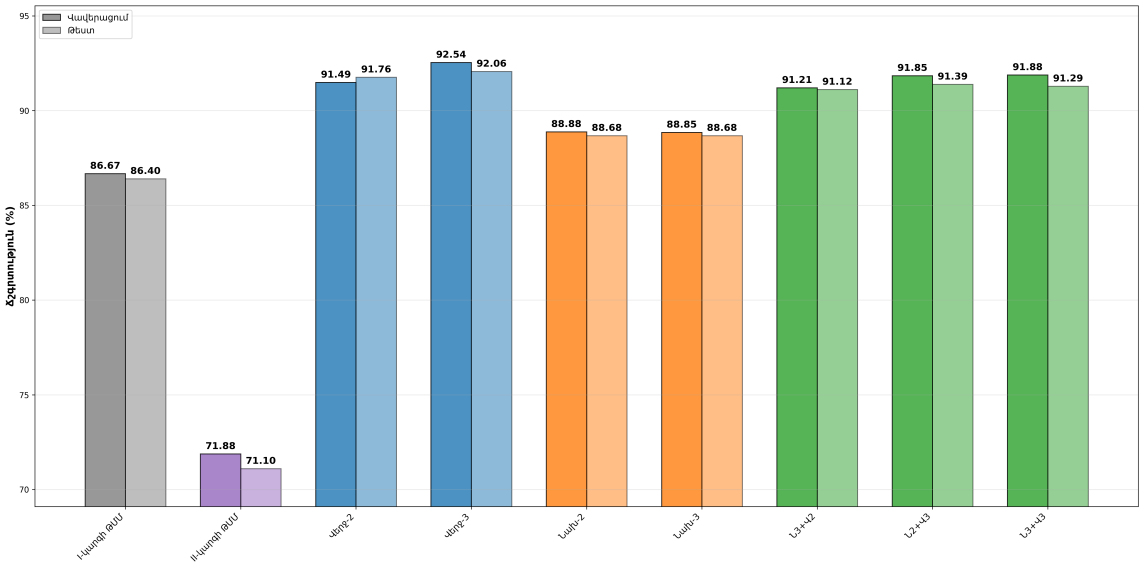
Ուսուցման համար նախատեսված տվյալների բազմությունը բաղկացած է 1974 նախադասությունից (42069 բառ), վավերացման բազմությունը՝ 249 նախադասությունից (5359 բառ), իսկ թեստային բազմությունը՝ 277 նախադասությունից (5157 բառ): Ընդհանուր առմամբ, տվյալների կորպուսը ունի 2500 նախադասություն և 52585 բառ: Տվյալների հավաքածուն հետևում է «Universal Dependencies» ստանդարտին՝ ունենալով 17 խոսքի մասերի պիտակներ: Դրանք են՝ NOUN (գոյական), VERB (բայ), ADJ (ածական), ADP (կապ), DET (դերանուն), PRON (դերանուն), AUX (օժանդակ բայ), PROPN (հատուկ անուն), ADV (մակբայ), CCONJ (համադասական շաղկապ), SCONJ (ստորադասական



Նկար 2: Խոսքի մասերի պիտակների բաշխումը ուսուցման տվյալների բազմությունում:

շաղկապ), PART (վերաբերական), NUM (թիվ), PUNCT (կետադրական նշան), INTJ (բացականչություն), SYM (սիմվոլ) և X (չպիտակավորված):

Այս հետազոտական աշխատանքի արդյունքում օգտագործվել է Վիտերբիի և պոսթերիոր վերծանման ալգորիթմները, թաքնված Մարկովյան մոդելով խոսքի մասերի պիտակավորման համար, ստանալով համապատասխանաբար 85.78% և 78.73% ճշգրտություն: Այս արդյունքի ցածր լինելը պայմանավորված է տվյալների կորպուսի փոքր լինելով, որը հանգեցնում է անծանոթ բառերի մեծ քանակին (27.15%) թեստային բազմությունում: Հետագա ուսումնասիրությունները ցույց տվեցին, որ եթե անծանոթ բառերի համար դիտարկվի դրանց վերջածանցը (վերջին 3 տառը) ուսուցման բազմության մեջ, և այդպես որոշվի դրանց ամենահավանական



Նկար 3: Մոդելների համեմատությունը Վիտերբիի ալգորիթմով: Ձախից աջ՝ սովորական թաքնված Մարկովյան մոդել, երկրորդ կարգի թաքնված Մարկովյան մոդել, անծանոթ բառերի համար վերջածանցների, նախածանցների, և միաժամանակ այդ երկուսի օգտագործումով:

Մոդել	Ճշգրտություն (%)
Առկա լավագույն մոդելը	81.25
Մեր մոդելը պոսթերիոր ալգորիթմով	78.73
Մեր մոդելը Վիտերբիի ալգորիթմով	85.78
Մեր մոդելը Վիտերբիի ալգորիթմով օգտագործելով վերջածանցը	92.06

Աղյուսակ 1: Թաքնված Մարկովյան մոդելների համեմատությունը:

պիտակը, ապա թաքնված Մարկովյան մոդելի ճշգրտությունը կլինի 92.06%, որը 10.81%-ով ավելի է, քան առկա ուսումնասիրություններում ստացված ամենալավ արդյունքը թաքնված Մարկովյան մոդելներով հայոց լեզվի խոսքի մասերի պիտակավորման համար: Ուսումնասիրությունները ցույց տվեցին, որ նշված արդյունքը ամենալավն է կատարված բոլոր փորձերի մեջ՝ օգտագործելով նաև նախածանցը, կամ նույնիսկ օգտագործելով երկրորդ կարգի Մարկովյան շղթայով $(P(q_i = x|q_1...q_{i-1}) = P(q_i = x|q_{i-1}, q_{i-2}))$ թաքնված Մարկովյան մոդել:

6 Եզրակացություն

Այս հետազոտական աշխատանքի ընթացքում ուսումնասիրությունները ցույց տվեցին, որ չնայած տվյալների կորպուսի փոքր և ոչ լիարժեք ճշգրիտ լինելուն, հայոց լեզվի համար խոսքի մասերի պիտակավորումը թաքնված Մարկովյան մոդելներով հնարավոր է իրականացնել ավելի քան 92% ճշգրտությամբ: Այս արդյունքը լավագույնն է առկա հետազոտություններում, որտեղ օգտագործվել են թաքնված Մարկովյան մոդելներ նշված խնդրի համար: Պարզվեց նաև, որ վերծանման համար օգտագործվող Վիտերբիի ալգորիթմը ավելի լավ ճշգրտություն ապահովեց,

համեմատած պոսթերիոր վերծանման հետ: Ինչպես նաև արդյունքների նշանակալի բարելավում եղավ անծանոթ բառերի պիտակավորման համար առաջարկված մոտեցումները կիրառելով:

Հղումներ

Akbik, A., Blythe, D., and Vollgraf, R. (2018). Contextual String Embeddings for Sequence Labeling. In Proceedings of the 27th International Conference on Computational Linguistics (COLING).

Aliwy, A. H., Radie, R. A., and Hamed, H. S. (2015). HMM Based POS Tagging System for 8 Different Languages and Several Tagsets. Engineering and Technology Journal, 33(2), 326-337.

Wang, P., Qian, Y., Soong, F. K., He, L., Zhao, H. (2015). Part-of-speech tagging with bidirectional long short-term memory recurrent neural network. arXiv preprint, 1510.06168.

Saidi, R., Jarray, F., and Mansour, M. (2021). A BERT based approach for Arabic POS tagging. In In International Work-Conference on Artificial Neural Networks. Cham: Springer International Publishing, (pp. 311-321)

Goyal, M., Joshi, S., and Kulkarni, R. (2019). Part-of-Speech Tagging for Marathi Text using HMM and CRF. In Proceedings of the International Conference on Applied and Theoretical Computing and Communication Technology (ICATCCT).

Pham B. (2020). Parts of Speech Tagging: Rule-Based.

Manning, C. D. and Schütze, H. (1999). Foundations of Statistical Natural Language Processing. MIT Press.

Baghdsaryan V. (2023). Comparative Analysis of Hidden Markov Model and Bidirectional Long Short-Term Memory for POS Tagging in Eastern Armenian. International Journal of Scientific Advances. ISSN: 2708-7972.

Bărbulescu A. and Morariu D. (2020). Part of Speech Tagging Using Hidden Markov Models. International Journal of Advanced Statistics and ITC for Economics and Life Sciences, 10(1).

Kumawat D. and Jain V. (2015). POS tagging approaches: A comparison. International Journal of Computer Applications, 118.6

Eddy, S. R. (2004). What is a hidden Markov model? Nature Biotechnology, 22(10), 1315–1316. doi:10.1038/nbt1004-1315

Jurafsky, D., and Martin, J. H. (2025). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models (3rd ed. draft). Unpublished manuscript. Draft available at: https://web.stanford.edu/ jurafsky/slp3/

Viterbi A.J. (1967). Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm. IEEE Transactions on Information Theory, 13(2), pp. 260–269.

Omura J.K. (1969). On the Viterbi Decoding Algorithm. IEEE Transactions on Information Theory, 15(1), pp. 177–179.

Forney G.D., Jr. (1974). Convolutional Codes II. Maximum-Likelihood Decoding. Information and Control, 25(3), pp. 222–266.

Gormley, M. (2023). *Lecture 20: Bayesian Networks*. 10-301/10-601 Introduction to Machine Learning, Carnegie Mellon University. Available at: https://www.cs.cmu.edu/~mgormley/courses/10601-s23//slides/lecture20-bayesnet.pdf.

Georgetown University. (2016). Viterbi Algorithm. Lecture slides. Available at: https://people.cs.georgetown.edu/cosc572/f16/12_viterbi_slides.pdf