Azərbaycan dili üçün söz vektorları və böyük dil modelləri

Ümid Süleymanov[1], Samir Rüstəmov[2], Abzətdin Adamov[3]
[1,2,3]ADA Universiteti, Bakı, Azərbaycan
[1]*usuleymanov@ada.edu.az*, [2]*srustamov@ada.edu.az*, [3]*aadamov@ada.edu.az*

***Xülasə* — Son illərdə söz vektorları və əvvəlcədən öyrədilmiş dil təmsil sistemlərinin uğuru təbii dilin işlənməsinə marağı xeyli artırmışdır. Onlar suallara cavab, tərcümə sistemləri, əhval-ruhiyyənin təhlili və s kimi Təbii Dilin Emalı tapşırıqlarında nəzərəçarpan irəliləmələrin əsasını təşkil edir. Söz vektorları üzərində tezis, metodların sürətlə artması nəzərə alınmaqla, az resurslu aqqlütinativ dil olan Azərbaycan dili üçün xüsusilə aktualdır. Bu tezis söz vektorlarının müxtəlif arxitektura yanaşmalarından istifadə edilməklə hazırlanmasını hədəfləyir və söz əlavələrinin effektivliyi hisslərin təhlili, mətn təsnifatı, semantik analogiya, sintaktik analogiya daxil olmaqla, müxtəlif təbii dil emal tapşırıqlarını ehtiva edən müxtəlif məlumat dəstləri üzərində empirik tədqiqatlar vasitəsilə araşdırılacaqdır.**

Açar sözlər — söz vektorları; sentiment analizi; təbii dilin emalı; statik vektorlar

I. GİRİŞ

İnsan intellekti ilə təbii dil arasında əlaqənin mövcudluğuna inanılması, təbii dil emalı sahəsini tətdqiqat üçün ən maraqlı sahələrindən birinə çevirir. Universal süni intellekt modellərini inkişaf etdirməzdən əvvəl, təbii dilin emalı ilə bağlı problemləri həll etmək çox vacibdir, çünki yalnız təbii dilin emalı ünsiyyət vasitəsilə, ağıllı sistemlər özlərini ən effektiv şəkildə ifadə edə bilərlər. Hətta dahi alim və kompüter elmlərinin banilərindən biri olan Alan Turinq özünün məşhur Turing testində [1] sorğu aparan şəxsin divarın arxasında maşınla, yoxsa insanla danışdığını müəyyən edə bilmirsə, maşının zəkaya malik olmasını təklif etmişdi. Bu, süni intellekt sahəsində təbii dilin emalı və generasiyasının əhəmiyyətinin daha bir sübutudur. Son zamanlardakı öncədən öyrədilmiş təbii dilin emalı modellərinin böyük uğurları da sahədəki tədqiqatı daha da zəruri edir.

Təbii dilin neyron təsvirində inqilab edən və onun müxtəlif sahələrində bir sıra irəliləyişlər yaradan söz vektorlarının yaradılması və istifadəsi bu tezisin əsas diqqət mərkəzində olacaqdır. Bu araşdırmanın məqsədi müxtəlif qabaqcıl metodlardan və maşın öyrənmə üsullarından istifadə edərək söz vektorları yaratmaqdır. Stabil söz vektorlarının yaradılması üçün publisistik, elmi və ədəbi üslubda yazılmış xəbər məqalələri, kitablar, elmi məqalələr, şeirlər və romanlar kimi müxtəlif mənbələrdən milyonlarla məlumat korpusu toplanmışdır. Söz vektorlarının effektivliyinin ölçülməsi üçün 2 yanaşma mövcuddur; intrinsik və extrinsik. İntrinsik yanaşmalarda [2] söz vektorlarının effektivliyi daxili analogiya tapşırığında ölçülür, ekstrinsik yanaşmada [3][4] isə

vektorlar hər hansı xarici təbii dilin emalı tapşırığında yoxlanılır.

II. ƏLAQƏLİ ƏDƏBİYYAT

Təbii dilin təsviri üsullarının müxtəlif aspektlərini əhatə edən zəngin ədəbiyyat mövcuddur. Collados və Pilehvar (2018) [5] müasir təbii dilin neyron şəbəkə üsullarından istifadə metodlarının xülasəsini təqdim edir və dil modelləşdirmə sistemlərinin sözün mənasını təmsil etməyə çalışarkən qarşılaşdığı çətinlikləri təqdim edir. Nəzarətsiz təlim vasitəsilə öyrənilən söz təmsilləri və nəzarətli təlimə əsaslanan təmsillər, texnikaları təmin etdikləri iki əsas kateqoriyadır. Collados və Pilehvar tərəfindən aparılan araşdırmada, həmçinin söz vektorlarının keyfiyyətinin ölçülməsi üçün qiymətləndirmə üsulları da müzakirə olunur (2018). Onlar tədqiqatı söz təmsillərinin istifadəsini təsvir etməklə və onları çoxlu faktorlar, o cümlədən mənanın kontekstə [6] uyğunluğu, şərh oluna bilməsi və s. müxtəlif kontekstlərdə tətbiqi baxımından qiymətləndirməklə yekunlaşdırırlar.

Mövcud dərin öyrənmə yanaşmaları və onların Təbii Dil Emalına təsirlərinin icmalı Otter və digərləri [7] tərəfindən özlərinin tədqiqat işlərində təqdim olunur. (2020). Təbii dilin emalının geniş icmalı və mövcud dərin öyrənmə arxitekturalarının araşdırılması işin birinci hissəsini təşkil edir. Təbii dilin emalı və dərin öyrənmə indi iki böyük mövzu olsa da, ayrı-ayrılıqda fərqli sahələr deyillər, onların kəsişməsini daha ətraflı araşdırmaq daha yaxşı olardı. Birinci hissədə dərin öyrənmə arxitekturasının izahı tez tamamlana bilərdi. Aşağıdakı bölmə dilin modelləşdirilməsi, morfologiyası, təhlili və semantikası daxil olmaqla [7] təbii dil emalının fundamental sahələrində dərin öyrənmənin istifadəsinə həsr edilmişdir (Otter və digərləri, 2020).

Peter və Pantel sözlərin semantik mənasının təmsil olunmasına diqqət yetirən əlavə bir araşdırma [8] aparmışlar (2010). Bu araşdırma ona görə fərqlənir ki, o, mövcud strukturları və metodları geniş təbii dil emal fərziyyələri ilə əlaqələndirməyə çalışır. Fərziyyələrin bir neçə nümunəsinə statistik semantik fərziyyə, sözlər paketi hipotezi, paylanma fərziyyəsi və başqaları daxildir. Nəticədə, sahəyə yeni daxil olan tədqiqatçılar bu araşdırmanı və onun araşdırdığı arxitekturaları başa düşməyi olduqca asan tapacaqlar. Tədqiqata linqvistik və riyazi baxımdan söz vektorlarının işlənməsinə dair ayrıca fəsillər də daxil edilmişdir.

III. METODOLOGİYA

Sözlərin yerləşdirilməsi üsulları son on il ərzində təbii dil emalı sahəsini kəskin şəkildə dəyişdirdi. Söz yerləşdirmələri müxtəlif maşın öyrənmə alqoritmlərini insanın təbii dilinin son dərəcə real təsvirini təmin edir. Tədqiqatçılar üçün əlçatan olan yaxşı formalaşmış korporasiyanın olmaması səbəbindən Azərbaycan dili aşağı resurs səviyyəsinə malikdir. Bundan əlavə, bu sintaktik və semantik cəhətdən mürəkkəb dilə yönəlmiş çox tədqiqat yoxdur. Azərbaycan dili aqqlütinativ dildir, yəni sözlər və qrammatik cümlələr, adətən prefiks və postfikslərdən istifadə etməklə əmələ gəlir. Təkcə bu fakt bu dil üçün söz yerləşdirmə konsepsiyasının effektivliyinin öyrənilməsini daha da münasib edir və bütün dünya üzrə təbii dil emalı

tədqiqatçıları və mütəxəssisləri üçün araşdırılması vacib bir mövzuya çevirir.

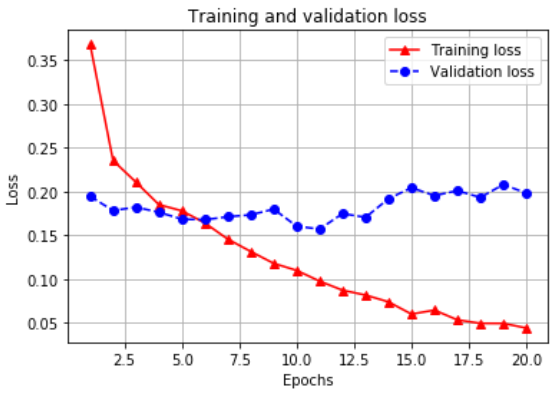
Sözlərin yerləşdirilməsi insan təbii dilinin sıx vektor təsvirlərini yaradır ki, bu da bir çox təbii dilin işlənməsi tapşırıqlarına, o cümlədən suallara cavab vermə, əhval-ruhiyyənin təhlili, mətnin təsnifatı və digərləri üçün uğurla tətbiq edilir və tez-tez, hətta bu tapşırıqlarda yeni ən müasir performansa nail olur. Bu, artıq danılmaz faktdır. Bununla belə, tədqiqatların əksəriyyəti ingilis, fransız, alman və başqaları kimi dil ailələri üzərində cəmləşib və bu kontekstlərdə söz əlavələrinin tapşırıqlar üzərində effektivliyini nümayiş etdirir. Bu da öz növbəsində aşağıdakı tədqiqat məsələsini ortaya qoyur: "Sözün yerləşdirilməsi yanaşmaları Azərbaycan dili kimi aqqlütinativ dili uğurla təmsil edirmi?"

Biz bu tədqiqat problemini həll etmək üçün həm daxili, həm də xarici qiymətləndirmə tapşırığı parametrlərində söz daxiletmələrinin nəticələrini kəmiyyətləndirməyə və qiymətləndirməyə çalışacağıq. Bunun üçün söz daxiletmələrindən istifadə edərək bir sıra təcrübələr aparılacaqdır. *Word2vec* alqoritminin tərtibatçıları Mikolov və digərləri [9] tərəfindən məşhur şəkildə təklif olunan bu tapşırıqlardan söz daxiletmələrinin effektivliyini statistik qiymətləndirmək üçün istifadə etmək geniş şəkildə qəbul edilir. Nəzərə alın ki, bu tapşırıqların ayrıca Azərbaycan dilində versiyası hazırlanmışdır. Bu daxili qiymətləndirmə tapşırıqlarında həm semantik, həm də sintaktik bənzətmə tapşırıqlarından istifadə edilir və sözün yerləşdirilməsinin düzgünlüyü qiymətləndirilir. Xarici qiymətləndirmə tapşırıqları üçün əhval-ruhiyyənin təhlili və mətn təsnifatı tapşırıqlarında sözlərin yerləşdirilməsinin effektivliyi göstəriləcək və dəqiqlik reytinqləri göstəriləcək. Təcrübələrdən sonra bu aqqlütinativ dil üçün söz əlavələrinin effektivliyi müzakirə ediləcək və verilən tədqiqat sualına empirik nəticələr tapılmağa çalışılacaq.

FastText arxitekturası N-qram simvollarından istifadə edərək, altsöz məlumatlarının uçotunu təmin edir. Nəzərə alsaq ki, Azərbaycan dili aqqlütinativ dildir, bu, xüsusilə faydalıdır. Ən çox sintaktik qarşılıqlı əlaqənin altsöz səviyyəsində baş verdiyini nəzərə alsaq aqqlütinativ dildə altsöz strukturlarına müraciət daha da vacib olur. Nəticədə, bu arxitekturanın altsöz səviyyəli qatları, sintaktik düzgünlüyün daxili qiymətləndirilməsini tələb edən tapşırıqlarda daha yaxşı çıxış edəcəyini təxmin edə bilərik. Üç fərqli məlumat dəstində öyrədilmiş müxtəlif ölçülü söz vektor ölçüləri ilə biz aşağıdakı cədvəldə fastText arxitekturası üçün hərtərəfli dəqiqlik reytinqlərini veririk. Ən yüksək ballar hər bir məlumat dəsti ölçüsü qruplaşması üçün qalın götürülür və altından xətt çəkilir. Modelin öyrəndiyi söz vektor ölçülərinin sayı vektor ölçüsü ilə təmsil olunur.

Söz yerləşdirmə arxitekturaları tərəfindən yaradılan söz vektorları hisslərin təhlili tapşırığına tətbiq edilir. Maşın öyrənmə arxitekturasının qiyməti söz vektorlarının keyfiyyətini əks etdirir. Sözün yerləşdirilməsi vektoru nə qədər yaxşı olarsa, model xarici qiymətləndirmə

tapşırığını bir o qədər yaxşı öyrənə bilər. Biz dondurulmuş yerləşdirmə vektorlarını sentiment təhlili tapşırığında yerləşdirmə təbəqəsi kimi istifadə etdik və maşın öyrənmə arxitekturamızı öyrətdik. Maşın öyrənmə arxitekturamız Yerləşdirmə Layeri, nizamlanma üçün Dropout qatı, 96 x 64 Convolution Layer, MaxPooling Layer, LSTM qatı və tam əlaqəli sıx təbəqədən ibarətdir. Arxitektura ümumilikdə 33 225 965 parametrə malikdir.



Şəkil 1. Öyrədilmə dəyərinin aşağı enməsi

Dövrlərin funksiyası kimi model arxitekturasının təlim və təsdiqləmə itkisi yuxarıdakı şəkildə göstərilmişdir. Göründüyü kimi, dövrlərin sayı artdıqca məşq itkisi azalır, lakin doğrulama itkisi 0,20 ətrafında statik nümunə nümayiş etdirir. Bu, modelin məlumat dəstindən kifayət qədər öyrəndiyi üçün 20 dövr sayının kifayət olduğunu göstərir. Beləliklə, daha çox dövrlər əlavə etmək həddindən artıq uyğunlaşma ilə nəticələnəcəkdir. Model təsdiqləmə məlumatlarını görmədiyi üçün təlim itkisindən daha böyük qiymətləndirmə itkisi adi haldır, buna görə də doğrulama itkisi qaçılmaz olaraq təlim itkisindən daha yüksəkdir. Model təlim dəstindəki nümunəni uğurla öyrəndi, bunu təsdiqləmə və təlim itkisi üçün yaxın dəyərlər sübut edir.

İSTİNADLAR

[1] Turing, A.M. (1950), 'Computing Machinery and Intelligence', Mind 59, pp. 433–460

[2] Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of NAACL-HLT 2019, pages 4171–4186.

[3] Mikolov, T., Yih, W.-t., & Zweig, G. (2013d). Linguistic regularities in continuous space word representations. In HLT-NAACL, pp. 746–751.

[4] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013c). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems, pp. 3111–3119.

[5] Jose Camacho-Collados & Mohammad Taher Pilehvar. (2018). From word to sense embeddings: a survey on vector representations of meaning. J. Artif. Int. Res. 63, 1 (September 2018), 743–788. DOI:https://doi.org/10.1613/jair.1.11259 [7] Joseph Turian, Lev Ratinov, and Yoshua Bengio. (2010). Word representations: a simple and general method for semi-supervised learning. In Proceedings of ACL, pages 384–394

[6] Melamud, O., Goldberger, J., & Dagan, I. (2016). context2vec: Learning generic context embedding with bidirectional lstm. In Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning, pp. 51–61, Berlin, Germany

[7] D. W. Otter, J. R. Medina and J. K. Kalita, (2020). "A Survey of the Usages of Deep Learning for Natural Language Processing," in IEEE Transactions on Neural Networks and

Learning Systems, doi: 10.1109/TNNLS.2020.2979670

[8] Turney Peter D. & Pantel Patrick (2010). From Frequency to Meaning: Vector Space Models of Semantics. Journal of Artificial Intelligence Research, 37, 141- 188.

[9] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013a). Efficient estimation of word representations in vector space. CoRR, abs/1301.3781.

[10] Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. (2018.) QANet: Combining local convolution with global self-attention for reading comprehension. In ICLR. 71

[11] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In International Conference on Machine Learning, pages 1188– 1196

[12] Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of NAACL-HLT 2019, pages 4171–4186.

[13] Qing Cui, Bin Gao, Jiang Bian, Siyu Qiu, Hanjun Dai, and Tie-Yan Liu. (2015). KNET: A general framework for learning word embedding using morphological knowledge. ACM Transactions on Information Systems, 34(1):4:1–4:25.

[14] Eric H. Huang, Richard Socher, Christopher D. Manning, and Andrew Y. Ng. (2012). Improving Word representations via global context and multiple word prototypes. In Proceedings of Association for Computational Linguistics, pages 873–882.

[15] John Duchi, Elad Hazan, and Yoram Singer. (2011). Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. J. Mach. Learn. Res. 12, (2/1/2011), 2121–2159

[16] [13] Melamud, O., McClosky, D., Siddharth Patwardhan, and Mohit Bansal. 2016. The role of context types and dimensionality in learning word embeddings. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1030-1040.

[17] Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In Proceedings of NAACL, New Orleans, LA, USA.