

**TABIY TILNI QAYTA ISHLASHDA SEMANTIK O'XSHASHLIKNI
ANIQLASHNING FORMAL USULLARI**

Abdullayev Elmurod Zaylobidinovich – *pedagogika fanlari bo'yicha falsafa doktori, dotsent*
elmurod_abdullayev@mail.ru
Andijon davlat universiteti

**ФОРМАЛЬНЫЕ МЕТОДЫ ОПРЕДЕЛЕНИЯ СЕМАНТИЧЕСКОГО ПОДОБИЯ
ПРИ ОБРАБОТКЕ ЕСТЕСТВЕННОГО ЯЗЫКА**

Абдуллаев Элмурод Зайлобидинович – *доктор философии по педагогическим наукам, доцент*
Андижанский государственный университет

**FORMAL METHODS FOR DETERMINING SEMANTIC SIMILARITY
IN NATURAL LANGUAGE PROCESSING**

Abdullayev Elmurod Zaylobidinovich – *Doctor of Philosophy in Pedagogical Sciences, associate professor*
Andijan State University

Tayanch so'zlar: tabiiy tilni qayta ishlash, semantik o'xshashlik, formal model, vektor fazosi, algoritmik tahlil.

Rezyume. Maqolada tabiiy tilni qayta ishlash jarayonida semantik o'xshashlikni aniqlash masalasi nazariy va formal yondashuvlar asosida tahlil qilinadi. Semantik o'xshashlik tushunchasi matematik va algoritmik modellar orqali aniqlanib, vektor fazosiga asoslangan, ehtimolli hamda grafik modellarning nazariy asoslari yoritiladi. Tadqiqotda semantik o'xshashlikni baholashning formal mezonlari ishlab chiqilib, ularning umumlashtirish qobiliyati va hisoblash xossalari bilan bog'liqligi asoslab beriladi. Olingan natijalar NLP algoritmlarini nazariy baholash va takomillashtirish uchun ilmiy asos bo'lib xizmat qiladi.

Ключевые слова: обработка естественного языка, семантическое сходство, формальные методы, векторное представление, алгоритмический анализ.

Резюме. В статье рассматриваются формальные методы определения семантического сходства в обработке естественного языка. Понятие семантического сходства анализируется с позиций математических и алгоритмических моделей, включая векторные, вероятностные и графовые подходы. Разрабатываются формальные критерии оценки семантического сходства и обосновывается их связь с обобщающей способностью и вычислительными свойствами алгоритмов NLP. Полученные результаты могут быть использованы для теоретической оценки и совершенствования алгоритмов обработки естественного языка.

Key words: natural language processing, semantic similarity, formal methods, vector space model, algorithmic analysis.

Summary. This paper examines formal methods for measuring semantic similarity in natural language processing. It discusses the concept of semantic similarity using mathematical and algorithmic models, such as vector-space, probabilistic, and graph-based approaches. Formal criteria for evaluating semantic similarity are proposed, and their links to generalization ability and computational features are scientifically supported. The findings offer a foundational framework for assessing and enhancing NLP algorithms.

Kirish. Tabiiy tilni qayta ishlash (NLP) sohasida semantik o'xshashlikni aniqlash muammosi markaziy o'rinlardan birini egallaydi. Matnlar, so'zlar yoki tushunchalar o'rtasidagi semantik yaqinlikni aniqlash axborot qidiruvi, matnlarni klasterlash, mashina tarjimasini, savol-javob tizimlari va bilimlarni chiqarish kabi ko'plab masalalarning nazariy asosini tashkil etadi. Shu sababli, semantik o'xshashlikni aniqlash usullarini formal modellar asosida o'rganish dolzarb ilmiy vazifa hisoblanadi.

Amaliy tadqiqotlarda semantik o'xshashlik ko'pincha empirik ko'rsatkichlar orqali baholanadi. Biroq, nazariy informatika nuqtayi nazaridan ushbu tushunchani matematik va algoritmik jihatdan asoslash, uning hisoblash xossalari aniqlash muhim ahamiyatga ega. Ayniqsa, NLP algoritmlarining umumlashtirish qobiliyati va barqarorligi semantik o'xshashlikni aniqlashning formal mezonlariga bevosita bog'liqdir.

Mazkur maqolaning maqsadi tabiiy tilni qayta ishlashda semantik o'xshashlikni aniqlashning formal usullarini nazariy jihatdan tahlil qilish, ularni matematik modellar asosida tasniflash va baholashdan iborat.

Mavzuga oid adabiyotlar tahlili. Semantik o'xshashlik tushunchasi dastlab tilshunoslik va kognitiv fanlarda

shakllangan bo'lsa-da, keyinchalik u informatika va mashinaviy o'rganish nazariyasida formal modellar orqali ifodalanadigan tushunchaga aylandi. Vektor fazosiga asoslangan modellarning nazariy asoslari Salton tomonidan axborot qidiruvi kontekstida ishlab chiqilgan [1]. Keyinchalik bu yondashuv NLP sohasiga keng tatbiq etildi.

Jurafsky va Martin ishlarida semantik o'xshashlikning statistik va ehtimolli modellar orqali ifodalanishi keng yoritilgan [2]. Word embedding modellarining nazariy asoslari Mikolov va boshqalar tomonidan ishlab chiqilgan bo'lib, ular semantik yaqinlikni vektor fazosidagi masofa orqali aniqlash imkonini berdi [3]. Goldberg esa neyron tarmoqlarga asoslangan semantik modellarni formal nuqtayi nazardan tahlil qilgan [4].

Biroq, mavjud tadqiqotlarning aksariyati amaliy tajribaga yo'naltirilgan bo'lib, semantik o'xshashlikning formal va algoritmik xossalari yetarli darajada tizimlashtirilmagan. Shu bois, mazkur maqolada ushbu bo'shliqni to'ldirishga harakat qilinadi.

Tadqiqot metodologiyasi. Tadqiqotda matematik modellash, vektor fazosi nazariyasi, ehtimollik va statistik tahlil, algoritmik murakkablikni baholash kabi nazariy metodlar qo'llanildi.

Semantik o'xshashlik $\text{sim}(x,y)$ funksiyasi sifatida aniqlanib, u so'z yoki tushunchalar orasidagi masofani formal tarzda ifodalaydi. Ushbu funksiya metrik yoki psevdometrik xossalarga ega bo'lishi mumkin.

Asosiy qism. Tabiiy tilni qayta ishlashda semantik o'xshashlikni aniqlash masalasi algoritmik va matematik modellashirish nuqtayi nazaridan murakkab ko'p omilli jarayon hisoblanadi. Semantik o'xshashlik til birliklarining ma'nodagi yaqinligini ifodalaydigan abstrakt tushuncha bo'lib, uni formal ifodalash uchun til birliklarini matematik obyektlar sifatida tasvirlash talab etiladi. Nazariy informatika doirasida bu masala avvalo, mos fazoni aniqlash va ushbu fazoda masofa yoki yaqinlik funksiyasini kiritish orqali hal etiladi [1].

Eng keng tarqalgan formal yondashuvlardan biri vektor fazosiga asoslangan modellar hisoblanadi. Ushbu modellar doirasida so'zlar, iboralar yoki matn birliklari R^n kabi yuqori o'lchamli vektor fazosida tasvirlanadi. Bunda har bir koordinata ma'lum statistik yoki semantik xususiyatni ifodalaydi. Semantik o'xshashlik esa ikki vektor orasidagi kosinus o'xshashlik, Evklid masofasi yoki boshqa metrik funksiyalar orqali aniqlanadi [1, 9]. Ushbu yondashuvning nazariy afzalligi shundaki, u chiziqli algebra va metrik fazolar nazariyasiga tayanadi hamda algoritmik jihatdan nisbatan samaralidir.

Biroq vektor fazosiga asoslangan modellar semantik munosabatlarning barcha jihatlarini to'liq ifodalay olmaydi. Xususan, sinonimiya va polisemantiklik kabi hodisalar ba'zi hollarda faqat statistik yaqinlik bilan chegaralanib qoladi. Shu sababli ehtimolli modellar semantik o'xshashlikni aniqlashda muhim alternativ yondashuv sifatida qaraladi. Bu yondashuvda semantik o'xshashlik til birliklarining kontekstual ehtimollik taqsimotlari orqali aniqlanadi [2]. Masalan, ikki so'zning bir xil yoki o'xshash kontekstlarda uchrash ehtimoli ularning semantik yaqinligini ifodalaydi. Ushbu yondashuv Bayes ehtimollik nazariyasiga va statistik o'rganish modeliga asoslanadi [6, 8].

Ehtimolli modellar semantik noaniqlikni hisobga olish imkonini bersa-da, ularning hisoblash murakkabligi yuqori bo'lishi mumkin. Shu bois, nazariy informatika nuqtayi nazaridan ehtimolli yondashuvlar algoritmik barqarorlik va umumlashtirish qobiliyati bilan bog'liq holda tahlil qilinadi [11]. Tadqiqotlar shuni ko'rsatadiki, ehtimolli modellar to'g'ri regularizatsiya qilinganda, semantik o'xshashlikni barqaror va ishonchli aniqlash imkonini beradi [6].

Semantik o'xshashlikni aniqlashning yana bir muhim formal yondashuvi grafik modellar bilan bog'liq. Ushbu yondashuvda til birliklari grafik tugunlar sifatida, ularning semantik munosabatlari esa qirralar sifatida ifodalanadi. Masalan, WordNet kabi leksik resurslarda sinonimiya, giperonimiya va antonimiya munosabatlari grafik tuzilma orqali ifodalanadi [5]. Semantik o'xshashlik grafikdagi eng qisqa yo'l uzunligi yoki tugunlar orasidagi strukturaviy yaqinlik orqali baholanadi. Bu yondashuv mantiqiy va ierarxik munosabatlarni formal tarzda ifodalash imkonini beradi.

Nazariy tahlil shuni ko'rsatadiki, grafik modellar diskret strukturalarga asoslangan bo'lib, ular semantik munosabatlarni tushuntiriladigan shaklda tasvirlaydi. Biroq, katta hajmdagi matnlar uchun grafik modellarni qurish va

qayta ishlash katta hisoblash resurslarini talab etadi [7]. Shu sababli, zamonaviy tadqiqotlarda grafik yondashuvlar ko'pincha vektor yoki ehtimolli modellar bilan integratsiyalashgan holda qo'llaniladi.

Umuman olganda, semantik o'xshashlikni aniqlashning formal usullari bir-birini inkor etmaydi, balki o'zaro to'ldiruvchi xususiyatga ega. Vektor modellar uzluksiz fazoda statistik yaqinlikni ifodalaydi, ehtimolli modellar noaniqlikni hisobga oladi, grafik modellar esa mantiqiy va strukturaviy munosabatlarni aniqlaydi [3, 4, 10]. Nazariy informatika nuqtayi nazaridan, ushbu yondashuvlarni yagona formal model doirasida umumlashtirish NLP algoritmilarining umumlashtirish qobiliyati va ishonchligini oshirishga xizmat qiladi.

Shu tariqa, semantik o'xshashlikni aniqlash masalasi NLP algoritmilarining fundamental nazariy xususiyatlaridan biri bo'lib, u algoritmik samaradorlik, hisoblash murakkabligi va barqarorlik bilan uzviy bog'liqdir. Ushbu masalani formal modellar asosida chuqur tahlil qilish kelajakda nazariy jihatdan asoslangan va tushuntiriladigan sun'iy intellekt tizimlarini yaratish uchun mustahkam ilmiy poydevor bo'lib xizmat qiladi [2, 7, 11].

Natijalar va muhokama. Nazariy tahlil natijalari shuni ko'rsatadiki, semantik o'xshashlikni aniqlashning formal usullari NLP algoritmilarining umumlashtirish qobiliyatiga bevosita ta'sir ko'rsatadi. Vektor fazosiga asoslangan modellar hisoblash nuqtayi nazaridan samarali bo'lsa-da, ular kontekstual va mantiqiy munosabatlarni to'liq aks ettira olmaydi. Ehtimolli modellar esa noaniqlikni hisobga olish imkonini bersa-da, ularning hisoblash murakkabligi yuqori bo'lishi mumkin.

Grafik modellarga asoslangan yondashuvlar semantik munosabatlarni aniq ifodalash imkonini beradi, biroq ular katta hajmdagi ma'lumotlar bilan ishlashda murakkablik muammosiga duch keladi. Shu bois, formal nuqtayi nazardan eng samarali yechim ushbu yondashuvlarni integratsiyalash orqali amalga oshirilishi mumkin.

Nazariy va formal tahlillar natijasida aniqlandiki, semantik o'xshashlikni aniqlash tabiiy tilni qayta ishlash tizimlarining samaradorligini belgilovchi eng muhim komponentlardan biri hisoblanadi. Tadqiqot davomida vektor fazosiga asoslangan modellar, ehtimolli yondashuvlar hamda grafik modellarning nazariy va algoritmik xususiyatlari chuqur tahlil qilindi. Olingan natijalar shuni ko'rsatadiki, semantik o'xshashlikni aniqlashda har bir model alohida afzalliklarga ega bo'lib, ular NLP algoritmilarining umumlashtirish qobiliyati va hisoblash samaradorligiga turlicha ta'sir ko'rsatadi.

Vektor fazosiga asoslangan modellar semantik yaqinlikni matematik jihatdan ifodalashda eng samarali formal yondashuvlardan biri sifatida namoyon bo'ldi. Ushbu modellar doirasida so'zlar va matn birliklari yuqori o'lchamli vektorlar ko'rinishida tasvirlanadi hamda ular orasidagi o'xshashlik kosinus o'xshashligi yoki masofa funksiyalari orqali aniqlanadi. Tadqiqot natijalariga ko'ra, mazkur yondashuv katta hajmdagi korpuslarni qayta ishlashda yuqori tezkorlikni ta'minlaydi va algoritmik jihatdan samarali hisoblanadi. Biroq ayrim hollarda statistik yaqinlik haqiqiy semantik yaqinlikni to'liq ifodalamaydi. Ayniqsa, polisemantik birliklar va kontekstga bog'liq

ma'nolarni aniqlashda vektor modellarining cheklovlari kuzatildi.

Ehtimolli modellarni nazariy tahlil qilish jarayonida ularning semantik noaniqlikni hisobga olish imkoniyatiga ega ekanligi aniqlandi. Ushbu yondashuvda til birliklarining kontekstual ehtimollik taqsimotlari asosiy mezon sifatida qaraladi. Natijalar shuni ko'rsatdiki, ehtimolli modellar murakkab semantik munosabatlarni aniqlashda nisbatan yuqori aniqlik beradi. Ayniqsa, kontekstual ma'no farqlarini aniqlashda Bayes nazariyasiga asoslangan modellar samarali natija ko'rsatadi. Shu bilan birga, ehtimolli modellar uchun katta hajmdagi ma'lumotlar va yuqori hisoblash resurslari talab qilinishi ularning amaliy qo'llanishini murakkablashtiradi.

Grafik modellarga asoslangan yondashuvlar esa semantik bog'lanishlarni strukturaviy va mantiqiy jihatdan ifodalash imkonini berdi. Tadqiqot davomida WordNet kabi leksik grafiklar asosida qurilgan modellar sinonimiya, antonimiya va giperonimiya munosabatlarini aniqlashda yuqori interpretatsiyalanuvchanlikka ega ekanligi aniqlandi. Grafik modellar semantik strukturalarni tushuntirish nuqtayi nazaridan samarali bo'lsa-da, katta hajmdagi tarmoqlar bilan ishlashda algoritmik murakkablik ortishi kuzatildi. Bu esa real vaqt rejimida ishlovchi NLP tizimlarida muhim muammo sifatida namoyon bo'ladi.

Muhokamalar natijasida semantik o'xshashlikni aniqlashda yagona modeldan foydalanish yetarli emasligi asoslandi. Chunki tabiiy til ko'p qatlamli murakkab tizim bo'lib, unda statistik, mantiqiy va kontekstual munosabatlar birgalikda shakllanadi. Shu sababli, vektor, ehtimolli va grafik modellarni integratsiyalashgan holda qo'llash eng samarali yondashuv sifatida baholandi. Integratsiyalashgan formal modellar NLP algoritmlarining barqarorligi, aniqligi va umumlashtirish qobiliyatini oshirish imkonini beradi. Olingan nazariy natijalar semantik o'xshashlikni formal modellar asosida chuqur o'rganish kelgusida tushuntiriladigan va ishonchli sun'iy intellekt tizimlarini yaratishda muhim ilmiy asos bo'lib xizmat qilishini ko'rsatdi.

Xulosa. Mazkur maqolada tabiiy tilni qayta ishlashda semantik o'xshashlikni aniqlash masalasi formal va nazariy informatika nuqtayi nazaridan tahlil qilindi. Semantik o'xshashlik tushunchasi matematik va algoritmik modellar orqali ifodalanib, uning NLP algoritmlarining umumlashtirish qobiliyati va hisoblash xossalari bilan uzviy bog'liqligi asoslab berildi.

Tadqiqot natijalari shuni ko'rsatdiki, semantik o'xshashlikni aniqlash uchun vektor fazosi, ehtimolli va grafik modellari asoslangan yondashuvlar har biri alohida nazariy afzalliklarga ega. Shu bilan birga, ushbu yondashuvlarning cheklovlari ham mavjud bo'lib, ularni kompleks qo'llash orqali bartaraf etish mumkin.

Tadqiqot natijalari asosida shuni ta'kidlash mumkinki, semantik o'xshashlikni aniqlashning formal usullari NLP sohasida nazariy va amaliy tadqiqotlarni birlashtiruvchi strategik yo'nalish hisoblanadi. Ayniqsa, formal modellashtirish orqali NLP algoritmlarining interpretatsiyalanuvchanligi, barqarorligi va umumlashtirish qobiliyatini oshirish imkoniyati paydo bo'ladi. Bu esa kelgusida sun'iy intellekt tizimlarining ishonchligini oshirish va inson tilini chuqurroq tushunadigan

algoritmarni yaratish uchun muhim metodologik asos bo'lib xizmat qiladi.

Xulosa qilib aytganda, semantik o'xshashlikni formal modellarga asoslangan holda o'rganish NLP sohasida fundamental ilmiy ahamiyatga ega bo'lib, bu yo'nalishdagi tadqiqotlar kelgusida ishonchli va tushuntiriladigan sun'iy intellekt tizimlarini yaratishga xizmat qiladi.

Takliflar:

1. Semantik o'xshashlikni baholash uchun yagona formal mezonlar tizimini ishlab chiqish. Hozirgi kunda tabiiy tilni qayta ishlashda semantik o'xshashlikni aniqlash turli empirik va kontekstga bog'liq mezonlar orqali amalga oshirilmoqda. Bu holat NLP algoritmlarini nazariy jihatdan solishtirish va ularning umumlashtirish qobiliyatini baholashda muhim qiyinchiliklarni keltirib chiqaradi. Shu sababli, semantik o'xshashlikni aniqlash uchun metrik va psevdometrik xossalarga ega bo'lgan yagona formal mezonlar tizimini ishlab chiqish maqsadga muvofiqdir. Bunday mezonlar vektor fazosi, ehtimollik taqsimotlari va grafik strukturalar asosida aniqlanib, NLP algoritmlarini nazariy baholash uchun umumiy standart vazifasini bajarishi mumkin.

2. Vektor, ehtimolli va grafik modellarni integratsiyalashgan nazariy model doirasida birlashtirish. Tadqiqot natijalari shuni ko'rsatdiki, semantik o'xshashlikni aniqlashda qo'llaniladigan mavjud formal yondashuvlarning har biri alohida afzallik va cheklovlarga ega. Vektor modellar hisoblash jihatdan samarali bo'lsa, ehtimolli modellar noaniqlikni hisobga oladi, grafik modellar esa mantiqiy va strukturaviy munosabatlarni aniqlik bilan ifodalaydi. Shu bois, ushbu yondashuvlarni yagona integratsiyalashgan nazariy model doirasida birlashtirish taklif etiladi. Bunday model semantik o'xshashlikni ko'p qatlamli formal struktura sifatida ifodalash imkonini berib, NLP algoritmlarining nazariy asoslarini yanada mustahkamlaydi.

3. Semantik o'xshashlikni aniqlash algoritmlarining hisoblash murakkabligini nazariy tahlil qilish. Semantik o'xshashlikni aniqlash algoritmlarining samaradorligi nafaqat aniqlik, balki ularning hisoblash murakkabligi bilan ham belgilanadi. Shu sababli, ushbu algoritmlarning vaqt va xotira murakkabligini formal modellar asosida chuqur tahlil qilish muhim hisoblanadi. Ayniqsa, katta hajmdagi matnlar va yuqori o'lchamli vektor fazolari bilan ishlash sharoitida algoritmik murakkablikning semantik aniqlikka ta'sirini nazariy jihatdan baholash dolzarbdir. Bu yo'nalishdagi tadqiqotlar optimallashtirilgan va barqaror NLP algoritmlarini yaratishga xizmat qiladi.

4. NLP tizimlarini loyihalashda formal semantik mezonlarni optimallashtirish kriteriyasi sifatida joriy etish. Amaliy NLP tizimlarini ishlab chiqishda ko'pincha empirik samaradorlik ko'rsatkichlari asosiy mezon sifatida qo'llaniladi. Biroq, tadqiqot natijalari shuni ko'rsatdiki, formal semantik mezonlarni optimallashtirish kriteriyasi sifatida joriy etish NLP algoritmlarining umumlashtirish qobiliyati va ishonchligini oshiradi. Shu bois, kelgusida NLP tizimlarini loyihalashda semantik o'xshashlikning formal mezonlarini algoritmik optimallashtirish jarayoniga kiritish tavsiya etiladi. Bu yondashuv nazariy jihatdan asoslangan, tushuntiriladigan va barqaror sun'iy intellekt tizimlarini yaratishga zamin yaratadi.

Adabiyotlar

1. Salton G. Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. Reading (MA): Addison-Wesley, 1989.
2. Jurafsky D., Martin J.H. Speech and Language Processing. 3rd ed. – Pearson Education, 2023.
3. Mikolov T., Chen K., Corrado G., Dean J. Distributed Representations of Words and Phrases and their Compositionality. // Advances in Neural Information Processing Systems. 2013. Vol. 26. – P. 3111-3119.
4. Goldberg Y. Neural Network Methods for Natural Language Processing. – San Rafael (CA): – Morgan & Claypool Publishers, 2017.
5. Fellbaum C. (Ed.) WordNet: An Electronic Lexical Database. – Cambridge (MA): MIT Press, 1998.
6. Vapnik V. N. Statistical Learning Theory. – New York: Wiley-Interscience, 1998.
7. Mohri M., Rostamizadeh A., Talwalkar A. Foundations of Machine Learning. 2nd ed. – Cambridge (MA): MIT Press, 2018.
8. Bishop C. M. Pattern Recognition and Machine Learning. – New York: Springer, 2006.
9. Manning C. D., Raghavan P., Schütze H. Introduction to Information Retrieval. – Cambridge: Cambridge University Press, 2008.
10. Cover T. M., Thomas J. A. Elements of Information Theory. 2nd ed. – Hoboken (NJ): Wiley-Interscience, 2006.
11. Bousquet O., Elisseeff A. Stability and Generalization. // Journal of Machine Learning Research. Vol. 2, 2002. – P. 499-526.
- Zhang T. Statistical Behavior and Consistency of Classification Methods. // Annals of Statistics. Vol. 32, №1, 2004. – P. 56-134.