Semantik ravishda olish uchun so'zlarni joylashtirishdan foydalanish Jamiyat savollariga javob berishda o'xshash savollar

Nuha Usmon, Rim Faiz, Kamel Smayli

Semantik ravishda olish uchun so'zlarni joylashtirishdan foydalanish Jamiyat savoliga o'xshash savollar Javob berish

NOUHA OTHMAN1, RIM FAIZ2, VA KAMEL SMAÏLI3

1LARODEC, Tunis universiteti - Tunis
2LARODEC, Karfagen universiteti - Tunis
3LORIA, Lotaringiya universiteti - Frantsiya
1othmannouha@gmail.com
2rim.faiz@ihec.rnu.tn
3smaili@loria.fr

2018 yil 23 fevralda tuzilgan

Ushbu maqolada asosiy e'tibor Hamjamiyatda hal qiluvchi va qiyin vazifa bo'lgan savollarni qidirishga qaratilgan Savolga javob berish (CQA). Savollarni qidirish tarixiy savollarni topishga qaratilgan so'ralganlarga semantik jihatdan teng, o'xshash savollarga javoblar deb faraz qilgan holda yangilariga ham javob berishi kerak. Asosiy muammolar leksik bo'shliq muammosi bilan bir qatorda tabiiy tilda so'zlashuv. Ko'pgina mavjud usullar o'rtasidagi o'xshashlikni o'lchaydi so'zlar sumkasi (BOW) ko'rinishiga asoslangan savollar, ular orasida hech qanday semantika yo'q so'zlar. Ushbu maqolada biz savollarning mazmunli vektor ko'rinishi uchun so'zlarni joylashtirish va TF-IDF-ga tayanamiz. Savollar orasidagi o'xshashlik kosinus o'xshashligi yordamida o'lchanadi vektorga asoslangan so'z ko'rinishlariga asoslanadi. Tajribalar haqiqiy dunyo ma'lumotlari bo'yicha o'tkazildi Yahoo'dan to'plam! Javoblar bizning usulimiz raqobatbardosh ekanligini ko'rsatadi.

1. KIRISH

Jamiyatga asoslangan savol-javoblar (CQA), bu taqdim etadi Turli xil ma'lumotlarga ega bo'lgan odamlarning bilim almashishi uchun platformalar tobora ommalashib borayotgan axborot vositasiga aylandi Internetda qidirish. CQA-da foydalanuvchilar o'zaro aloqada bo'lishlari va javob berishlari mumkin boshqa foydalanuvchilarning savollari yoki boshqa ishtirokchilar javob berishlari uchun o'z savollarini joylashtiring [1]. So'nggi yillarda, Internetning bumi bilan 2.0, cQA Yahoo! kabi foydalanuvchi tomonidan yaratilgan kontentni ishlab chiqarish uchun onlayn xizmatning qiziqarli shakli sifatida paydo bo'ladi Javoblar 1 Stackover-oqim2, MathOverflow3 , LinuxQuestions4 va boshqalar. Bunday jamoat xizmatlari katta savol-javob arxivlarini yaratdi doimiy ravishda ko'payib borayotgan juftliklar takrorlanadi savollar. Shunday qilib, foydalanuvchilar yuzlab mumkin bo'lgan javoblar orasidan osonlikcha to'g'ri javoblarni topa olmaydi va keyin yangisini joylashtiradi arxivlarda allaqachon mavjud bo'lgan savollar. kamaytirish uchun yangi javob olish uchun vaqt kechikishi kerak, cQA avtomatik ravishda kerak ekvivalent savollar mavjudligini tekshirish uchun jamoa arxivini qidiring

1http://answers.yahoo.com/

2http://stackoverflow.com/
3http://www.mathoverflow.net
4http://www.linuxquestions.org/

ilgari e'lon qilingan. Agar shunga o'xshash savol aniqlansa, unga tegishli javob to'g'ridan-to'g'ri yangi so'rovga tegishli javob sifatida qaytarilishi mumkin .

So'nggi paytlarda ushbu yo'nalish bo'yicha ko'plab qiziqarli tadqiqotlar o'tkazildi [2-7] yangi savollarga o'tmishdagi javoblar bilan javob berish uchun. Darhaqiqat, savollarni izlash bir nechta qiyinchiliklarga duch keladigan ahamiyatsiz vazifadir, chunki cQA savollari lug'at, uzunlik, uslub va kontent sifati jihatidan sezilarli darajada farq qiladi. Eng katta qiyinchilik so'ralayotgan savollar va arxivdagi mavjudlar o'rtasidagi leksik bo'shliqdir [2], bu an'anaviy ma'lumot qidirish (IR) modellari uchun haqiqiy to'siqdir, chunki foydalanuvchilar bir xil savolni turli xil iboralar yordamida shakllantirishlari mumkin. Masalan, savollar: Sizning ishingizning xususiyatlari qanday? va Ishingizni qanday tasvirlay olasiz?, bir xil ma'noga ega, lekin ular leksik jihatdan farq qiladi. So'zning nomuvofiqligi CQAda juda muhim muammodir, chunki savollar nisbatan qisqa va shunga o'xshash savollar odatda siyrak va so'zlarning bir-biriga mos kelmasligi bilan ifodalanadi . Bundan ko'rinib turibdiki, keng ko'lamli jamoat arxivlaridan to'liq foydalanish uchun savollarni qidirish uchun samarali qidiruv modellari juda zarur . CQAdagi leksik bo'shliq muammosini bartaraf etish uchun ko'pgina zamonaviy tadqiqotlar savollar o'rtasidagi o'xshashlik o'lchovini yaxshilashga harakat qiladi, shu bilan birga so'zlarning siyrak va diskret ko'rinishlari uchun o'xshashlik funktsiyasini o'rnatish qiyin . Eng muhimi, mavjud yondashuvlarning aksariyati kontekstual ma'lumotni hisobga olmaydi va so'zlar o'rtasidagi etarlicha semantik aloqalarni qamrab olmaydi.

So'zlarni joylashtirish deb ham ataladigan taqsimlangan semantik tasvirlarni o'rganish bo'yicha so'nggi sa'y-harakatlar tabiiy tilni qayta ishlash (NLP) va so'z o'xshashligi [8], tavsiya tizimlari [9] va savollarni qidirish [10] kabi ko'plab IR vazifalari uchun ajoyib imkoniyat ekanligi ko'rsatildi . So'zlarni joylashtirish - bu lug'at tarkibidagi so'zlarni past o'lchamli (lug'at hajmi bilan taqqoslaganda) haqiqiy vektorlarga solishtirishga qaratilgan yangi texnika . Ushbu bo'shliqda yaqin vektorlar mos keladigan so'zlar orasidagi yuqori semantik va sintaktik o'xshashlikni ko'rsatishi kerak . Garchi so'zlarni o'rnatish ko'plab qiyin vazifalarda sezilarli samaradorlikni ko'rsatgan bo'lsa-da , savollarni qidirish vazifasini yaxshilash uchun so'zlarni o'rnatishdan foydalanish haqida kam ma'lumot mavjud.

Ushbu yangi paydo bo'lgan usullarning so'nggi muvaffaqiyatidan kelib chiqib, ushbu maqolada biz CQAda savollarni qidirish uchun so'zlarni joylashtirishga asoslangan usulni taklif qilamiz . Savollarni so‘zlar to‘plami (BoW) sifatida ifodalash o‘rniga , so‘zlarni o‘rnatishning eng mashhur modeli bo‘lgan word2vec yordamida ularni uzluksiz bo‘shliqda o‘rnatilgan so‘zlar sumkasi (BoEW) sifatida ko‘rsatishni taklif qilamiz. So'zlarning semantik so'z birikmalaridan foydalangan holda ifodalanishi savollardagi semantik ma'lumotlarning ko'p qismini tushunishi kerak. Savolning yaratilgan so'z birikmalari so'ngra TF-IDF (terminal chastota - teskari hujjat chastotasi) ma'lumotlaridan foydalanish orqali o'lchanadi va savolning umumiy ko'rinishini olish uchun o'rtacha hisoblanadi. Qizig'i shundaki, TF-IDF vazni bilan birga so'zlarni ifodalash uchun so'zlarni joylashtirishdan foydalanish qisqa matn fragmenti uchun samarali vektor tasvirini topishda va'da berdi [11]. Shuning uchun savollar har bir savol uchun vektorga asoslangan so'z ko'rinishiga asoslangan kosinus o'xshashligidan foydalangan holda tartiblanadi . Oldingi qo'yilgan savol so'ralayotgan savolga semantik jihatdan o'xshash deb hisoblanadi, agar ularning tegishli vektor ko'rinishlari kosinus o'xshashligiga ko'ra bir-biriga yaqin bo'lsa.

o'lchov. Kosinus o'xshashligi eng yuqori ballga ega bo'lgan oldingi savol yangi e'lon qilingan savolga eng o'xshash savol sifatida qaytariladi . Biz taklif qilingan usulni Yahoo! Javoblar. Eksperimental natijalar shuni ko'rsatadiki, bizning usulimiz istiqbolli va cQAda savollarni qidirishning ba'zi zamonaviy usullaridan ustun bo'lishi mumkin .

Ushbu maqolaning qolgan qismi quyidagicha tashkil etilgan: (2) bo'limda biz CQAda savollarni qidirish bo'yicha asosiy bog'liq ishlarning umumiy ko'rinishini beramiz. So'ngra, biz (3) bo'limda biz taklif qilingan so'zlarni joylashtirishga asoslangan savolni qidirish usulini tasvirlaymiz. (4) bo'lim eksperimental baholashimizni taqdim etadi va (5) bo'lim maqolani yakunlaydi va ba'zi istiqbollarni belgilaydi.

2. ALOQALI ISH

So'nggi paytlarda jamoatchilik savollariga javob berish (CQA) xizmatlarining gullab-yashnashi bilan birga , CQAda savollarni qidirishga qiziqish ortib bormoqda . Semantik jihatdan o'xshash savollarni aniqlash uchun katta tadqiqot ishlari olib borildi , ularga bir xil javob bilan javob berish mumkin.

Bir nechta ishlar so'rov va arxivlangan savollar o'rtasidagi kosinus o'xshashligini hisoblash uchun VSM deb ataladigan vektor fazo modeliga asoslangan edi [3, 12]. Biroq, VSM ning asosiy cheklovi shundaki, u qisqa savollarni qo'llab-quvvatlaydi, cQA xizmatlari esa ixcham yoki faktoid savollar bilan cheklanmagan holda keng ko'lamli savollarni hal qila oladi . VSM ning kamchiligini bartaraf etish uchun BM25 savol uzunligini hisobga olgan holda savollarni qidirish uchun ishlatilgan [3]. Okapi BM25 - Robertson va boshqalar tomonidan taklif qilingan Okapi qidirish modellari oilasi orasida eng keng tarqalgan qo'llaniladigan model. da [13] va bir nechta IR vazifalarida sezilarli ishlashni isbotladi . Bundan tashqari, til modellari (LM) [14] so'rovlarni atamalar to'plami o'rniga so'rov shartlari ketma-ketligi sifatida aniq modellashtirish uchun ham ishlatilgan . LMlar atamalarning nisbiy pozitsiyalarini hisobga olgan holda har bir mumkin bo'lgan voris atama uchun nisbiy ehtimollikni baholaydi. Shunga qaramay, foydalanuvchi so'rovi va arxivlangan savollar o'rtasida umumiy so'zlar kam bo'lsa, bunday modellar samarali bo'lmasligi mumkin.

LMlar duch keladigan lug'at mos kelmasligi muammosini bartaraf etish uchun tarjima modeli parallel korpusga asoslangan so'zlar o'rtasidagi korrelyatsiyani o'rganish uchun ishlatilgan va u savollarni qidirishda sezilarli natijalarga erishgan. Tarjima modellari ortidagi asosiy sezgi savol-javob juftligini parallel matnlar sifatida ko'rib chiqishdan iborat bo'lib , so'zlarning munosabatini so'zdan so'zga tarjima qilish ehtimolini o'rganish orqali qurish mumkin, masalan [2, 4].

Xuddi shu kontekstda [15] statistik so'zlarni tarjima qilish modellarini o'rgatish uchun turli leksik semantik manbalar tomonidan bir xil atama uchun berilgan ta'riflar va glosslardan iborat parallel ma'lumotlar to'plamini taqdim etdi . [16] da mualliflar alohida soÿzlarni tarjima qilish oÿrniga , iboralar tarjimasini bir butun sifatida qurishda baÿzi kontekstual maÿlumotlarni qoÿshib, soÿzga asoslangan tarjima modelini yaxshilashga harakat qilganlar. [5] da soÿzga asoslangan tarjima modeli semantik maÿlumotlarni (obÿektlar) oÿz ichiga olgan holda kengaytirilgan va cQA arxivlari va obÿyektlar katalogidan foydalangan holda soÿzlar va tushunchalar oÿrtasidagi tarjima ehtimolini oÿrganish strategiyalarini oÿrgangan . Yuqorida aytib o'tilgan asosiy modellar yaxshi natijalar bergan bo'lsa-da , savol va javoblar aslida parallel emas, balki ular tarkibidagi ma'lumotlardan farq qiladi [6].

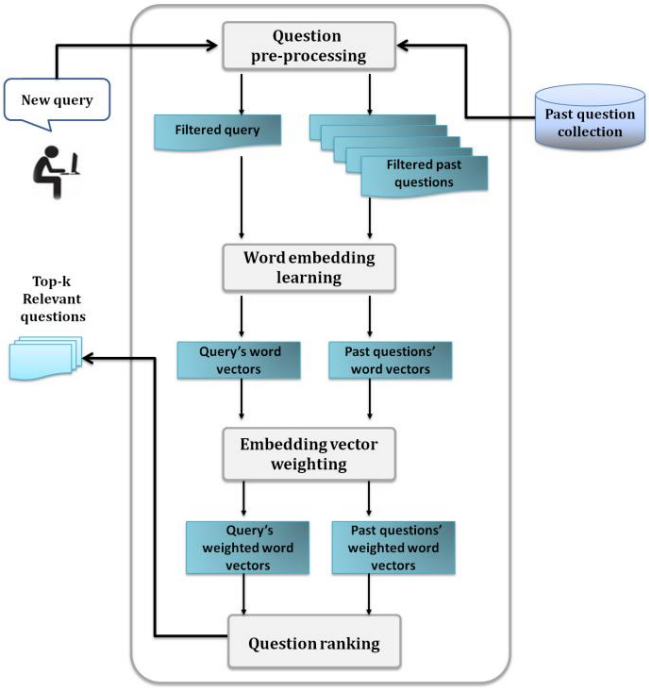
Semantik o'xshashlikka asoslangan ilg'or yondashuvlar ekvivalent savollarni aniqlash uchun qisqa matnni chuqur anglash yo'lida savoldagi leksik bo'shliq muammosini bartaraf etish uchun talab qilindi . Masalan, [3, 14, 17] kabi savollarni izlash uchun mavjud kategoriya maÿlumotlaridan foydalanishga urinishlar kam boÿlgan. Ushbu urinishlar savollarni izlash uchun til modelining ish faoliyatini sezilarli darajada yaxshilashini isbotlaganiga qaramay , toifa ma'lumotlaridan foydalanish faqat til modeli bilan cheklandi. Vang va boshqalar [18] savollarning sintaktik daraxtlarini yaratish uchun tahlilchidan foydalangan va ularni sintaktik daraxtlari va so'rov savoli o'rtasidagi o'xshashlik asosida tartiblagan . Shunga qaramay, bunday yondashuv juda murakkab, chunki u juda ko'p ta'lim ma'lumotlarini talab qiladi. [18] kuzatganidek , mavjud tahlilchilar norasmiy yozilgan savollarni tahlil qilish uchun hali ham yaxshi o'rganilmagan .

Bundan tashqari, o'tmishda izlangan savollar va nomzodlar o'rtasidagi semantik munosabatlarni chuqur savol tahlili bilan modellashtirishga ko'p urinishlar qilingan, masalan , savol mavzusini aniqlash va savollarni qidirishga e'tibor berishni taklif qilgan [12] . Shu nuqtai nazardan, ba'zi tadqiqotlar [19] kabi bir nechta xususiyatlar bilan olingan savollarni tartiblash yondashuvini taqdim etgan va [20] nomzodning javoblarini turli xususiyatlar kombinatsiyasi o'rniga bitta so'z ma'lumoti bilan tartiblagan. Yashirin semantik indekslash (LSI) [21] [22] dagi kabi berilgan vazifani hal qilish uchun ham ishlatilgan . Bir xil kontseptsiya haqidagi so'zlarni bir-biri bilan taqqoslash orqali sinonimiya va polisemiyani hal qilish samarali bo'lsa-da, LSI samaradorligi ko'p jihatdan ma'lumotlar tuzilishiga bog'liq va uni o'rgatish va xulosa chiqarish katta lug'atlarda hisoblash qimmat.

Aks holda, boshqa ishlar past o'lchamli vektor fazoda so'zlarning taqsimlangan ko'rinishlarini o'rganish uchun paydo bo'lgan modelga tayangan holda, savollar uchun taqdimotni o'rganishga qaratilgan , ya'ni Word Embedding. Bu ikkinchisi yaqinda katta qiziqish uyg'otdi va ko'plab NLP vazifalarida [23, 24], xususan, savollarni qidirishda [10] va'da berdi. Ushbu nazoratsiz ta'lim modelining asosiy afzalligi shundaki, u qimmat izohga muhtoj emas; u faqat o'qitish bosqichida juda katta miqdordagi xom matn ma'lumotlarini talab qiladi. So'zlarni ifodalash savolni qidirish vazifasi uchun juda muhim va oxirgi modelning muvaffaqiyatidan ilhomlanar ekan, biz cQAda savollarni qidirish vazifasini hal qilish uchun so'zlarni joylashtirishga tayanamiz.

3. WECOSIM TA'RIFI

WECOSim deb nomlangan savolni izlash uchun biz taklif qilayotgan usulning orqasidagi sezgi jamiyat to'plamidagi har bir savoldagi so'zlarni uzluksiz vektorlarga aylantirishdir . Har bir savolni so'zlar sumkasi (BOW) sifatida ifodalovchi an'anaviy usullardan farqli o'laroq, biz savolni o'rnatilgan so'zlar sumkasi (BoEW) sifatida taqdim etishni taklif qilamiz. Uzluksiz so'z ko'rinishlari uzluksiz so'zlar sumkasi (CBOW) modeli yordamida oldindan o'rganiladi [25]. Shuning uchun har bir savol uzluksiz bo'shliqqa kiritilgan so'zlar to'plami sifatida belgilanishi mumkin . Savolning so'z birikmalari TF-IDF ma'lumotlaridan foydalanish orqali o'lchanadi va savolning umumiy ko'rinishini olish uchun o'rtacha hisoblanadi . Bundan tashqari, kosinus o'xshashligi so'z vektorlarining o'rtacha qiymati o'rtasidagi o'xshashlikni hisoblash uchun ishlatiladi. Bilimlar so'ralgan savolga va arxivdagi har bir mavjud savolga . So'ngra tarixiy savollar kosinus o'xshashlik ballari bo'yicha saralanadi va yangi so'ralgan savolga eng mos bo'lgan maksimal ballga ega bo'lgan yuqori o'rindagi savolni qaytarish uchun. 1- rasmda ko'rsatilganidek , CQAda savollarni qidirish uchun taklif qilingan usul to'rt bosqichdan iborat: savollarga oldindan ishlov berish, so'zlarni joylashtirishni o'rganish, o'rnatish vektorini tortish va savollarni tartiblash.



1-rasm. Taklif etilayotgan usulning umumiy ko'rinishi

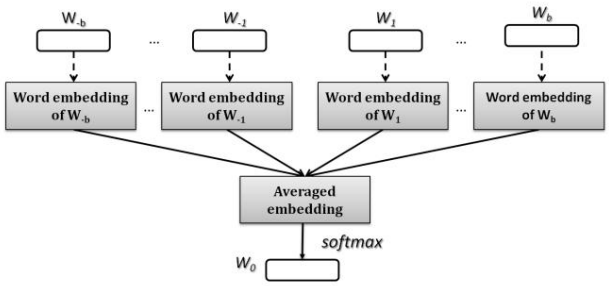
A. Savolga oldindan ishlov

berish Savolni oldindan qayta ishlash moduli rasmiy so'rovlarni yaratish uchun tabiiy tildagi savollarni qayta ishlash va foydali atamalarni ajratib olishni ko'zlaydi. Bular matnni tozalash, tokenizatsiya, to'xtash so'zlarni olib tashlash va stemmingni qo'llash orqali o Shunday qilib, savolni oldindan qayta ishlash moduli oxirida biz filtrlangan so'rovlar to'plamini olamiz, ularning har biri rasmiy ravishda quyidagicha aniqlanadi : q = t1 , t2, ..., tQ bu erda t so'rovning alohida atamasini ifodalaydi q va Q so'rov shartlari sonini bildiradi.

B. So'zni joylashtirishni

o'rganish Taqsimlangan semantik ko'rinishlar deb ham ataladigan so'zlarni joylashtirish usullari katta korpusdagi kontekstlari asosida uzluksiz so'z vektorlarini yaratishda muhim rol o'ynaydi . Ular har bir lug'at atamasi uchun past o'lchamli vektorni o'rganadilar , unda so'z vektorlari orasidagi o'xshashlik mos keladigan so'zlar orasidagi sintaktik va semantik o'xshashlikni ko'rsatishi mumkin. Asosan, so'zlarni joylashtirishning ikkita asosiy turi mavjud: Continuous Word Bag-of-Words modeli (CBoW) va Skip-gram modeli. Birinchisi hozirgi so'zni oldindan aytishdan iborat

kontekst, ikkinchisi esa suriluvchi oynada maqsadli so'z berilgan kontekstual so'zlarni teskari bashorat qiladi. Shuni ta'kidlash joizki , biz ushbu ishda so'zlarni joylashtirishni o'rganish uchun CBOW modelini [25] ko'rib chiqamiz , chunki u Skip-gramga qaraganda samaraliroq va katta hajmdagi ma'lumotlar bilan yaxshiroq ishlaydi. 2- rasmda ko'rsatilganidek , CBOW modeli kontekstning uzluksiz taqsimlangan so'zli ko'rinishidan foydalanib, atrofdagi so'zlarning ifodalanishini hisobga olgan holda markaziy so'zni bashorat qiladi, shuning uchun CBOW nomi. Kontekst vektori o'rtacha qiymatni olish orqali olinadi



2-rasm. Uzluksiz so'z xaltasi modelining umumiy ko'rinishi.

Har bir kontekstli so‘zning o‘rnatilishi, w0 markaziy so‘zining bashorati esa V lug‘atiga softmax qo‘llash orqali olinadi. Rasmiy ravishda, d so‘zni o‘rnatish o‘lchovi- |V|×d sion bo‘lsin, O ÿ chiqish matritsasi c kontekst vektorini | V|-o‘lchovli vektorga quyidagi markaz so‘zni ifodalovchi va prognozli vektorni ko‘rsatadi:

$$p(v0 \mid w[\ddot{y}b,b]\ddot{y}\{0\}) = \frac{\text{expv } 0^{T} \text{ ok}}{\ddot{y}v\ddot{y}V \text{ expv TOc}} \qquad (1)$$

Bu yerda b - kontekstli so'zlar oynasini belgilovchi giperparametr, Oc kontekst vektor c ning V lug'atdagi proyeksiyasini , v esa bir-issiq tasvirdir. CBOW ning kuchi shundaki, biz oynani kattalashtirganda u sezilarli darajada ko'tarilmaydi b.

C. Vektor vaznini o'rnatish. Savollar

o'rnatilgan so'zlar sumkasi (BoEW) sifatida taqdim etilgandan so'ng, yaratilgan vektorlar soddaligi va samaradorligi tufayli ma'lumot qidirish tizimlarida eng ko'p qo'llaniladigan tortishish sxemalaridan biri bo'lgan TF-IDF yordamida tortiladi.

Boshqacha qilib aytganda, har bir o'rnatilgan so'z o'zi ifodalagan so'zning TF-IDF ga ko'paytiriladi. TF-IDF - bu ma'lum bir hujjatdagi nisbiy chastotasi va butun hujjatlar to'plamidagi so'zni o'z ichiga olgan hujjatlarning teskari nisbati asosida so'zning ahamiyatini hisoblash uchun statistik tortish funktsiyasi . Savollar ustida ishlaganimizda, hujjatlarni oddiygina savollar bilan almashtirib, asosiy funksiyani kontekstimizga moslashtiramiz.

Savollar to‘plami C a so‘z w va q savolini hisobga olgan holda TF-IDF quyidagicha aniqlanadi:

$$\text{tf id f}(w, q, C) = \text{t f}(w, q)\ddot{y}\text{id f}(w, Q) = \text{fw,q}\ddot{y} \log(\ ) \text{ fw,C} \quad \frac{|C|}{} \qquad (2)$$

Bu yerda fw,q - q, |C| savolida wning necha marta paydo bo'lishi savollar to'plamining o'lchami va fw,C - jami w so'zini o'z ichiga olgan savollar soni . Biz TF-IDF dan so'zning nafaqat ma'lum bir savolda, balki butun savollar to'plamida qanchalik muhimligini baholash uchun foydalanamiz. Darhaqiqat, ba'zi umumiy so'zlar savollarda bir necha marta uchraydi, ammo ular indekslanadigan yoki qidiriladigan asosiy tushunchalar sifatida ahamiyatli emas.
Intuitiv ravishda, bitta yoki kichik savollar to'plamida keng tarqalgan so'zlarga yuqori ball beriladi, savollarda tez-tez uchraydigan so'zlar esa past ballga ega bo'ladi.

D. Savollar reytingi.

So'rov so'zlarining og'irlashtirilgan o'rnatilgan vektorlari so'ralgan savolning o'rtacha Vq vektorini quyidagi tarzda hisoblash uchun qo'llaniladi:

$$\text{Vq} = \frac{\text{V}|\ddot{y}_{i\ddot{y}1} \text{ (ywi} \times \text{tf id f(wi , q, C))}}{\text{V}|\ddot{y}_{i\ddot{y}1} \text{ tf identifikatori f(wi , q, C)}} \qquad (3)$$

bu yerda ywi — word2vec va |V| tomonidan yaratilgan wi soÿzining oÿrnatish vektori - berilgan savoldagi so'z vektorlari soni q . Xuddi shunday, har bir tarixiy savol uchun biz uning o'rtacha vektori Vd ni hisoblaymiz . So'ralgan savol bilan vektor fazodagi tarixiy savol o'rtasidagi o'xshashlik Vq va Vd o'rtasidagi kosinus o'xshashligi sifatida hisoblanadi . Savollar eng yuqori ballga ega bo‘lgan yangi so‘ralgan savolga eng mos bo‘lgan savollarni qaytarish uchun ularning vaznli vektorlari asosida kosinus o‘xshashlik ballari yordamida tartiblanadi .

4. TAJRIBALAR

A. Ma'lumotlar to'plami

Tajribalarimizda biz baholash uchun [26] tomonidan chiqarilgan ma'lumotlar to'plamidan foydalandik . Ma'lumotlar to'plamini yaratish uchun mualliflar Yahoo!-dagi barcha toifadagi savollarni o'rganib chiqdilar! Javoblar, eng mashhur cQA platformasi va keyin tasodifiy ravishda savollarni ikkita to'plamga bo'lib, ularning taqsimlanishini barcha toifalar bo'yicha saqlab qoldi. Birinchi to'plamda savollarni qidirish uchun savollar ombori sifatida 1,123,034 ta savol mavjud, ikkinchisi esa test to'plami sifatida ishlatiladi va 252 ta so'rov va 1624 ta qo'lda belgilangan tegishli savollarni o'z ichiga oladi. Har bir asl so'rovga tegishli savollar soni 2 dan 30 tagacha o'zgarib turadi. Savollar turli uzunlikdagi ikki so'zdan 15 so'zgacha, turli tuzilmalarda va turli toifalarga tegishli, masalan, Kompyuterlar va Internet , Yahoo! Mahsulotlar, oÿyin-kulgi va musiqa, taÿlim va maÿlumot, biznes va moliya, uy hayvonlari, sogÿlik, sport, sayohat, parhez va fitnes. 1- jadvalda test to'plamidagi so'rov va unga tegishli savollarning namunasi ko'rsatilgan. Annotatorlar

Jadval 1. Test to'plamidan savollarga misol.

Savol: Qanday qilib dietaga o'tirmasdan ozg'in bo'lishim mumkin?	
Kategoriya: Parhez va fitnes	
Mavzu: Og'irlikni yo'qotish **bilan**	
bog'liq - Men dietani o'zgartirmasdan qanday qilib sog'lom bo'laman? **Savollar** - Qanday qilib ozg'in bo'lishim mumkin, lekin na parhez, na jismoniy mashqlar?	- Qanday qilib parhez tabletkalarisiz tez oriq bo'lasiz? - Menga sog'lom bo'lish (vazn yo'qotish) uchun yechim kerak va aytishim kerakki, men qattiq dietalarni qabul qila olmaymanmi?

Agar nomzod savol semantik jihatdan so‘rovga o‘xshash yoki “ahamiyatsiz” deb hisoblansa, har bir so‘rovni “tegishli” bilan belgilash so‘ralgan . Agar ziddiyat yuzaga kelsa, uchinchi sharhlovchi yakuniy natijani baholaydi. E'tibor bering, test ma'lumotlaridagi savollar qidiruv ma'lumotlaridagi savollarga mos kelmaydi. So'zni o'rnatishni o'rgatish uchun biz cQA saytlaridan, ya'ni Yahoo! Webscope maÿlumotlar toÿplami5 , shu jumladan 2.512.345 ta soÿzdan iborat 1.256.173 ta savol. Tajribalardan oldin ba'zi dastlabki ishlov berish amalga oshirildi; barcha savollar kichik harflar bilan yozilgan, tokenlashtirilgan, Porter Stemmer6 tomonidan kelib chiqqan va barcha to'xtash so'zlari olib tashlandi.

B. So'zlarni joylashtirishni

o'rganish Biz butun Yahoo! Ta'lim ma'lumotlaridagi so'zlarni so'zlar kontekstini qamrab oladigan uzluksiz vektorlar sifatida ko'rsatish uchun word2vec-dan foydalangan holda Webscope ma'lumotlar to'plami. Word2vec ning o'quv parametrlari bir nechta testlardan so'ng o'rnatildi: xususiyat vektorlarining o'lchami 300 (hajmi = 300), kontekst oynasining o'lchami 10 (oyna = 10) va salbiy namunalar soni 25 (salbiy = 25) ga o'rnatildi.

C. Baholash ko'rsatkichlari

Usulimizning samaradorligini baholash uchun biz o'rtacha o'rtacha aniqlik (MAP) va Precision@n (P@n) dan foydalandik, chunki ular cQA uchun savollarni qidirish samaradorligini baholash uchun keng qo'llaniladi . Xususan, MAP adabiyotda eng ko'p qo'llaniladigan ko'rsatkich bo'lib , foydalanuvchi har bir so'rov uchun ko'plab tegishli savollarni topishdan manfaatdor deb hisoblaydi. MAP nafaqat tegishli savollarga erta javob beradigan, balki natijalarning yaxshi reytingini oladigan usullarni mukofotlaydi. So'ralgan Q savollari to'plamini hisobga olgan holda , MAP har bir so'ralgan q savolining o'rtacha aniqligini ifodalaydi va u quyidagicha o'rnatiladi:

$$MAP = \frac{ÿqÿQ\ AvgP(q)}{|\ Q\ |} \quad (4)$$

Bu erda AvgP(q) har bir tegishli savol q olingandan keyin aniqlik ballarining o'rtacha ko'rsatkichidir va u quyidagicha aniqlanadi:

$$O'rtacha\ P = \frac{ÿr\ P@r}{R} \quad (5)$$

Bu erda r har bir tegishli savolning darajasi, R - tegishli savollarning umumiy soni va P@r - topilgan r savollarning aniqligi .

Precision@n tegishli bo'lgan eng yuqori n-chi olingan savollarning nisbatini qaytaradi . So'ralgan Q savollari to'plamini hisobga olgan holda , P@n - so'rovlarga tegishli bo'lgan eng yuqori n ta olingan savollarning nisbati va u quyidagicha aniqlanadi:

$$P@n = \frac{1}{|\ Q\ |} \ddot{y}_{qÿQ} \frac{Nr}{N} \quad (6)$$

Bu erda Nr - q so'rovi uchun qaytarilgan eng yuqori N o'rinli ro'yxatdagi tegishli savollar soni . Tajribalarimizda biz P@10 va P@5 hisoblab chiqdik.

5Yahoo! Webscope maÿlumotlar toÿplami Yahoo keng qamrovli savollarga javob beradi va “http://research.yahoo.com/Academic_Relations” manzilida mavjud boÿlgan 1.0.2 versiyasiga javob beradi.

6http://tartarus.org/martin/PorterStemmer/

D. Asosiy natijalar

Biz WECOSim ish faoliyatini Zhang va boshqalar tomonidan sinovdan o'tgan quyidagi raqobatbardosh zamonaviy savollarni qidirish modellari bilan solishtiramiz. Xuddi shu ma'lumotlar to'plamida [26] :

- **TLM** [2]: Savol va javob qismidan foydalangan holda taxmin qilingan til modelini birlashtirgan tarjimaga asoslangan til modeli.

  U turli ma'lumot manbalaridan foydalangan holda o'rganilgan so'zdan so'zga tarjima qilish imkoniyatlarini birlashtiradi .

- **ETLM** [5]: ob'ektlarga semantik ma'lumotni kiritish uchun so'z tarjimasini ob'ekt tarjimasi bilan almashtirish orqali TLMning kengaytmasi bo'lgan ob'ektga asoslangan tarjima tili modeli .

- **PBTM** [16]: iboraga asoslangan tarjima modeli, u mashina tarjimasi ehtimolini qo'llaydi va savolni qidirish iboralar darajasida bajarilishi kerak deb taxmin qiladi.
  TLM tarixiy savoldagi so'zlar ketma-ketligini so'ralgan savoldagi boshqa so'zlar qatoriga tarjima qilish ehtimolini o'rganadi .

- **WKM** [27]: Og'irlik atamasi bahosini reyting funksiyasiga qo'shish uchun tashqi manba sifatida Vikipediyadan foydalangan dunyo bilimiga asoslangan model . Vikipediya haqidagi dunyo bilimlaridan olingan semantik munosabatlar asosida the-saurus kontseptsiyasi yaratilgan.

- **M-NET** [10]: Bir xil turkumga kiruvchi so'zlarning ko'rinishlari bir-biriga yaqin bo'lishi kerak, deb hisoblagan holda, yangilangan so'z o'rni olish uchun savollar toifalari ma'lumotlarini birlashtirgan doimiy so'zlarni joylashtirishga asoslangan model.

- **ParaKCM** [26]: pivot tillarning tarjimalarini o'rganuvchi va so'rovlarni parafrazalar bilan kengaytiruvchi asosiy tushunchani parafrazaga asoslangan yondashuv . Bu parafrazlar so'ralgan savoldagi asosiy tushunchalar va tarixiy savollar o'rtasida qo'shimcha semantik bog'lanishga hissa qo'shadi deb taxmin qiladi .

2- jadvaldan biz PBTM TLM dan ustun ekanligini ko'rishimiz mumkin, bu esa iboralarni butun yoki ketma-ket so'zlar ketma-ketligi sifatida tarjima qilishni modellashtirishda kontekstual ma'lumotni olish bir so'zni alohida-alohida tarjima qilishdan ko'ra samaraliroq ekanligini ko'rsatadi. Buning sababi, umuman olganda, iboradagi qo'shni so'zlar o'rtasida bog'liqlik mavjud. ETLM (an

2-jadval. Turli modellarning savollarni qidirish samaradorligini taqqoslash .

	TLM	ETLM	PBTM	WKM	M-NET	ParaKCM	WECOSim-tfidf	WECOSim	
P@5	0,3238	0,3314	0,3318	0,3413	0,3686	0,3722	**0,3432**	**0,4339**	
P@10	0,2548	0,2603	0,2603	0,2715	0,2848	0,2889	**0,2738**	**0,3646**	
MAP	0,3957	0,4073	0,4095	0,4116	0,4507	0,4578	**0,4125**	**0,5038**	

TLM kengaytmasi) PBTM kabi yaxshi ishlaydi, so'z tarjimasini reyting uchun ob'ekt tarjimasi bo'yicha qayta joylashtirishni isbotlaydi

tarjima tili modelining ishlashini yaxshilaydi. Garchi ETLM va WKM ikkalasi ham tashqi ma'lumot manbasiga asoslangan bo'lsa-da, masalan, Vikipediya, WKM bilim manbasidan kengroq ma'lumotdan foydalanadi . Xususan, WKM Vikipediyadagi tuzilmaviy bilimlar asosida tushuncha munosabatlarini (masalan, sinonimiya, gipernymiya, polisemiya va assotsiativ munosabatlar) keltirib chiqaradigan Vikipediya the-saurusini yaratadi. Tezaurusdagi turli munosabatlar so'rovni kengaytirish va keyin savolni qidirish uchun an'anaviy o'xshashlik o'lchovini kuchaytirish uchun ularning ahamiyatiga qarab ko'rib chiqiladi. Shunga qaramay, WKM va ETLM ning ishlashi turli foydalanuvchilarning savollariga Vikipediya tushunchalarining kam yoritilishi bilan cheklangan. M-NET, shuningdek, uzluksiz so'zlarni joylashtirishga asoslangan so'zlarning xususiyatlarini kodlash uchun toifa ma'lumotlarining metama'lumotlaridan foydalanish tufayli yaxshi ishlaydi , ulardan o'xshash so'zlarni ularning toifalari bo'yicha guruhlash mumkin. Eng yaxshi taqqoslangan tizim ParaKCM bo'lib, asosiy kontseptsiyaga asoslangan yondashuv bo'lib, u pivot tillarning tarjimalarini o'rganadi va savollarni qidirish uchun yaratilgan parafrazlar bilan so'rovlarni kengaytiradi.

Natijalar shuni ko'rsatadiki, bizning WECOSim usulimiz barcha mezonlar bo'yicha yuqorida aytib o'tilgan barcha usullardan sezilarli darajada ustunroq bo'lib, ko'plab tegishli savollarni erta topilganlar orasida qaytadan aylantiradi. Buning mumkin bo'lgan sababi shundaki, word2vec tomonidan o'rganilgan kontekst-vektor ko'rinishlari so'zlar o'rtasidagi semantik munosabatlarni aniqlash orqali so'zning leksik bo'shliq muammosini samarali hal qilishi mumkin, boshqa usullar esa semantik ekvivalentlik haqida etarli ma'lumotni to'plamaydi. Aytishimiz mumkinki, ko'milgan so'zlar bilan ifodalangan savollar, na semantikani, na matndagi pozitsiyalarni qamrab ololmaydigan an'anaviy so'zlar sumkasi modellariga qaraganda aniqroq yozilishi mumkin. Bu yaxshi ko'rsatkich TF-IDF og'irlik va kosinus o'xshashligi bilan birga so'zlarni joylashtirishdan foydalanish savolni qidirish vazifasida samarali ekanligini ko'rsatadi. Biroq, ba'zida bizning usulimiz shunga o'xshash savollarni ololmasligini aniqlaymiz : 252 ta test savolidan faqat 12 tasi P@10 qiymatini nolga tenglashtiradi. Ushbu savollarning aksariyati noto'g'ri yozilgan so'rov shartlarini o'z ichiga oladi. Misol uchun, dasturiy ta'minot atamasini o'z ichiga olgan so'rov uchun xatolik bilan dasturiy ta'minotni o'z ichiga olgan savollarni olib bo'lmaydi . Bunday holatlar shuni ko'rsatadiki, bizning yondashuvimiz ba'zi bir leksik kelishmovchilik muammolarini hal qila olmaydi. Bundan tashqari, WECOSim semantik ekvivalentlikni aniqlay olmaydigan holatlar kam . Ushbu holatlarning ba'zilari bitta o'xshash savolga ega bo'lgan savollarni o'z ichiga oladi va bu so'zlarning ko'pchiligi so'ralgan savolga o'xshash kontekstda ko'rinmaydi, masalan: Mening to'pimni qaysi tomonga qaratganim yaxshiroq, qutb yoki teshikmi ? va golfda qanday qilib maqsadga erishishim mumkin? Shubhasiz, natijalarni yaxshilash uchun qo'shimchalar o'lchamlari bilan qo'shimcha tajribalar o'tkazish kerak .

Boshqa tomondan, biz TF-IDF og'irliklari bilan va unsiz usulimizni sinovdan o'tkazdik (2-jadvalda WECOSim va WECOSim-tfidf mos ravishda) uning savollarni qidirish natijalariga ta'sirini o'rganish uchun. Tajribalarimiz orqali biz TF-IDF dan foydalanish P@5, P@10 va MAP qiymatlarini biroz oshirishga imkon berishini aniqladik. Buning sababi shundaki, TF-IDF aniq so'zlarni tez-tez ishlatadigan savollarni aniqlay oladi va ularning savolga mos kelishini aniqlaydi . Aytishimiz mumkinki, TF-IDF ning kamsituvchi kuchi qidiruv tizimiga yangi so'rovga o'xshash bo'lishi mumkin bo'lgan tegishli savollarni topishga imkon beradi. Biroq, ba'zi hollarda so'z nisbatan keng tarqalgan bo'lishi mumkin butun to'plam, lekin hali ham savol davomida sana va tizim so'zlari kabi muhim ahamiyatga ega . Bunday keng tarqalgan so'zlar past TF-IDF ball oladi va shuning uchun qidiruvda deyarli e'tiborga olinmaydi . Bundan tashqari, TF-IDF atamalar o'rtasidagi sinonimiya munosabatlarini hisobga olmaydi . Misol uchun, agar foydalanuvchi turar joy so'zini o'z ichiga olgan savolni joylashtirsa, TF-IDF so'rovga o'xshash bo'lishi mumkin bo'lgan savollarni ko'rib chiqmaydi, balki bungalov so'zini ishlatadi. TF-IDF bir xil tushuncha turli yo'llar bilan ifodalanishi mumkin bo'lgan norasmiy va heterojen savollar to'plamida tez-tez uchraydigan leksik noaniqlikni hal qila olmaydi . Shuni ham aytib o'tish joizki, TF-IDF ning hisoblash murakkabligi O(nm), bu erda n - so'zlarning umumiy soni va m - korpusdagi savollarning umumiy soni . Sizniki kabi katta to'plamlar uchun bu muammoni keltirib chiqarishi mumkin.

5. XULOSA

Bizning ishimiz CQA ning inson tomonidan yaratilgan mazmun ruhiga va o'tgan savol va javoblardan qayta foydalanishga to'g'ri keladi . Biz savolni qidirish vazifasiga e'tibor qaratamiz , shunga o'xshash o'tmishdagi savolga javoblar yangi savol ehtiyojlariga javob berishi kerak deb hisoblaymiz. Ushbu maqolada biz CQA arxivlaridan leksik bo'shliq muammosini hal qilish uchun so'zlarni joylashtirishga asoslangan usulni taklif qilamiz. Xususan, biz savollarni ifodalash uchun doimiy bo'shliqqa so'zlarni kiritishni taklif qilamiz . So'zlarni joylashtirish CBOW modeli yordamida oldindan o'rganiladi va so'zlarning chastotasiga qarab tortiladi.

Yangi so'rovga semantik jihatdan o'xshash savollarni topish uchun tarixiy savollar uzluksiz fazoda vektorga asoslangan so'z ko'rinishlariga asoslangan kosinus o'xshashligidan foydalangan holda tartiblanadi. Katta miqyosdagi cQA ma'lumotlari bo'yicha o'tkazilgan tajribalar savol so'zlarini ifodalash uchun TF-IDF bilan bir qatorda semantik so'zlarni joylashtirishdan foydalanish samaradorligini ko'rsatadi. Bizning usulimiz bir nechta umumiy so'zlarga ega bo'lsa ham, shunga o'xshash savollarni topishda mavjud bo'lganlardan sezilarli darajada ustun turishi mumkin . Biz TF-IDF vazni oddiy bo'lsa-da, qidiruv samaradorligi va qidiruv natijalari sifatini oshirishi mumkinligini isbotladik . Shunga qaramay, leksik noaniqlikni hisobga olmasdan, so'zni bitta vektor sifatida ko'rsatishning chegarasi mavjud . Kelgusi tadqiqotlar uchun biz oddiy TF-IDF sxemasini takomillashtirish va so‘z tasvirlarini boyitish maqsadida o‘quv jarayoniga har xil turdagi metama’lumotlar ma’lumotlarini kiritishni ko‘rib chiqamiz . Savollar orasidagi semantik o'xshashlikni hisoblash uchun turli xil o'xshashlik o'lchovlaridan foydalanishni o'rganish ham qiziqarli bo'ladi .

ADABIYOTLAR

1. Y. Liu, J. Bian va E. Agichteyn, " Jamoa savollariga javob berishda ma'lumot izlovchining qoniqishini bashorat qilish", " Axborot qidirishda tadqiqot va ishlanmalar bo'yicha 31-yillik ACM SIGIR konferentsiyasi materiallari " (ACM, 2008), 483-490-betlar.
2. X. Xue, J. Jeon va WB Croft, "Savol va javoblar arxivlari uchun qidiruv modellari ", " ACM SIGIR tadqiqot va tadqiqot bo'yicha 31-xalqaro konferentsiya materiallari" kitobida.

axborot qidirishda rivojlanish” (ACM, 2008), 475–482-betlar.

3. X. Cao, G. Cong, B. Cui va CS Jensen, “Umumiylashtirilgan savol uchun kategoriya ma'lumotlarini o'rganish doirasi hamjamiyat savollariga javoblar arxividan qidirish”, “ Jahon boÿyicha 19-xalqaro konferensiya maÿruzalari” kitobida. Wide Web”, (ACM, 2010), 201–210-betlar.
4. L. Cai, G. Chjou, K. Liu va J. Zhao, "Yashirin narsalarni o'rganish. jamiyatda savol izlash uchun mavzular qa.” “ Tabiat bo'yicha 5-Xalqaro qo'shma konferentsiya materiallari Tilni qayta ishlash”, (2011), 273–281-betlar.
5. A. Singx, “Shaxslarga asoslangan savol-javoblarni qidirish”, “Proceedings 2012-yilda tabiiy tilni qayta ishlash va tabiiy tilni hisoblashda empirik usullar boÿyicha qoÿshma konferensiya” (ACL, 2012), 1266–1277-betlar.
6. K. Chjan, V. Vu, X. Vu, Z. Li va M. Chjou, “Savol hamjamiyat savoliga yuqori sifatli javoblar bilan qidirish javob berish”, “23rd ACM International to‘g‘risidagi ma’lumotlar Axborot va bilimlar bo'yicha konferentsiya Menejment”, (ACM, 2014), 371–380-betlar.
7. P. Nakov, D. Xugeven, L. Markes, A. Moschitti, X. Muborak, T. Bolduin va K. Verspur, “Semeval-2017 3-topshiriq: Hamjamiyat savollariga javob berish”, “Tadbirlar Semantik baholash bo'yicha 11-xalqaro seminar (SemEval-2017),” (2017), 27–48-betlar.
8. T. Mikolov, I. Sutskever, K. Chen, GS Korrado va J. Din, “So‘z va iboralarning taqsimlangan ko‘rinishlari va ularning kompozitsionlik”, “Neyron axborotni qayta ishlash tizimlaridagi yutuqlar”, (2013), 3111–3119-betlar.
9. C. Musto, G. Semeraro, M. de Gemmis va P. Lops, “ Kontentga asoslangan soÿzlarni vikipediyadan oÿrganish. tavsiya qiluvchi tizimlar”, “Axborot qidirish boÿyicha Yevropa konferensiyasi ” (Springer, 2016), 729–734-betlar.
10. G. Chjou, T. Xe, J. Chjao va P. Xu, “Uzluksiz ta’lim savolni qidirish uchun metama'lumotlar bilan so'zni joylashtirish Jamiyat savoliga javob berish”, “Proceedings of the Hisoblash assotsiatsiyasining 53-yillik yig'ilishi Tilshunoslik va 7-Xalqaro qoÿshma konferensiya Osiyo Federatsiyasining tabiiy tillarni qayta ishlash Tabiiy tilni qayta ishlash”, (2015), 250–259-betlar.
11. C. De Boom, S. Van Canneyt, T. Demeester va B. Dhoedt, “Og'irlangan holda juda qisqa matnlar uchun vakillik o'rganish so‘zni o‘rnatish agregatsiyasi,” Pattern Recognit. Lett. **80,** 150–156 (2016).
12. H. Duan, Y. Cao, C.-Y. Lin va Y. Yu, “Qidiruv savollari savol mavzusini va savolga diqqat markazini aniqlash orqali. "ACL" da , jild. 8 (2008), jild. 8, 156–164-betlar.
13. SE Robertson, S. Walker, S. Jones, MM Hancock-Beaulieu, M. Gatford va boshqalar, "Okapi at trec-3", Nist Special Publ. Sp **109,** 109 (1995).
14. X. Cao, G. Kong, B. Cui, CS Jensen va C. Zhang, “The uchun til modellarida turkumlash ma'lumotlaridan foydalanish savollarni qidirish”, “ Axborot va bilimlarni boshqarish bo'yicha 18-AKM konferentsiyasi materiallari ” (ACM, 2009), 265–274-betlar.
15. D. Bernhard va I. Gurevych, “Leksik semantikani birlashtirish. Tarjima asosidagi javoblarni topish uchun savol-javob arxivlari bilan manbalar ”, “AKL va 47-yillik yig‘ilishining qo‘shma konferensiyasi materiallari”da. Tabiiy til bo'yicha 4-xalqaro qo'shma konferentsiya AFNLP ni qayta ishlash” (ACL, 2009), 728–736-betlar.
16. G. Chjou, L. Cai, J. Chjao va K. Liu, “ Jamoa savolida savollarni izlash uchun iboraga asoslangan tarjima modeli. javoblar arxivi”, “ Hisoblash tilshunosligi assotsiatsiyasining 49-yillik majlisi materiallari : Inson tili texnologiyalari-1-jild”, (ACL, 2011), s. 653–662.
17. G. Chjou, Y. Chen, D. Zeng va J. Chjao, “Tezroq sari. va savollarni qidirish uchun yaxshiroq qidirish modellari”, “ Axborot va bilimlarni boshqarish bo'yicha konferentsiya bo'yicha 22-ACM xalqaro konferentsiyasi materiallari” (ACM, 2013), 2139–2148-betlar.
18. K. Vang, Z. Ming va T.-S. Chua, “Jamoatga asoslangan qa xizmatlarida oÿxshash savollarni topish uchun sintaktik daraxtga mos yondashuv”, “ Axborot qidirishda tadqiqot va ishlanmalar boÿyicha 32-xalqaro ACM SIGIR konferentsiyasi materiallari ” (ACM, 2009), 187–194-betlar.
19. M. Surdeanu, M. Ciaramita va X. Saragosa, “O'rganish. katta onlayn qa to'plamlarida javoblarni tartiblash. "ACL" da, jild. 8 (2008), jild. 8, 719–727-betlar.
20. B. Vang, X. Vang, C. Sun, B. Liu va L. Sun, “Modellashtirish. ijtimoiy tarmoqdagi savol-javob juftliklari uchun semantik ahamiyatga ega jamoalar”, “48-yillik yig‘ilish materiallari Hisoblash tilshunosligi assotsiatsiyasi" (ACL, 2010), 1230–1238-betlar.
21. S. Deerwester, ST Dumais, GW Furnas, TK Landauer, va R. Xarshman, "Yashirin semantik tahlil orqali indekslash" J. Am. axborot fanlari jamiyati **41,** 391 (1990).
22. X. Qiu, L. Tyan va X. Huang, “Yashirin semantik tensor. jamiyatga asoslangan savollarga javob berish uchun indekslash. ichida “ACL (2),” (2013), 434–439-betlar.
23. J. Turian, L. Ratinov va Y. Bengio, “So‘zning ifodalanishi: “Yarim nazorat ostida o‘rganishning oddiy va umumiy usuli” bo‘limida “48-yillik yig‘ilish materiallari. Hisoblash tilshunosligi assotsiatsiyasi" (ACL, 2010), pp. 384–394.
24. R. Kollober, J. Veston, L. Bottou, M. Karlen, K. Kavukcuoglu, P. Kuksa, “Tabiiy til qayta ishlash (deyarli) noldan”, J. Mach. Learn. Res. **12,** 2493–2537 (2011).
25. T. Mikolov, K. Chen, G. Korrado va J. Din, “Samarali. vektor fazoda so'zlarning ko'rinishini baholash" arXiv oldindan chop etish arXiv: 1301.3781 (2013).
26. V.-N. Chjan, Z.-Y. Ming, Y. Zhang, T. Liu va T.-S. Chua, “Bir nechta kalit so'z birikmalarining semantikasini olish savollarni qidirish uchun tillar,” IEEE Transactions on Bilim. Data Eng. **28,** 888–900 (2016).
27. G. Chjou, Y. Liu, F. Liu, D. Zeng va J. Zhao, “Yaxshilash. yordamida jamiyatda savol-javobni qidirish dunyo bilimi”. "IJCAI", 13-jild (2013), jild. 2239–2245.

Nouha Usmon Tunis universitetining Institut Supérieur de Gestion (ISG) kompyuter fanlari bo'yicha bakalavr darajasini va Fransiyaning Nant universitetining ISG va Polytech Nantes kompyuter fanlari bo'yicha ikkita magistrlik darajasini oldi.

Hozirda u fan nomzodi. ISG Tunis, Tunis universitetida kompyuter fanlari nomzodi , 2015-yildan beri LARODEC laboratoriyasiga tegishli. Uning ilmiy qiziqishlari maÿlumot qidirish, tabiiy tillarni qayta ishlash va mashinani oÿrganishni oÿz ichiga oladi.

Rim Faiz doktorlik dissertatsiyasini oldi. hamkorlikda Parij-Dofin universiteti , LAMSADE laboratoriyasi, Frantsiyadagi fan.
Hozirda u Tunisdagi Karfagen universiteti LARODEC laboratoriyasining Oliy biznesni o'rganish instituti (IHEC) kompyuter fanlari bo'yicha professori. Uning ilmiy qiziqishlari axborot qidirish, katta ma'lumotlar, matn qazib olish, mashinada o'rganish, tabiiy tillarni qayta ishlash va semantik vebni o'z ichiga oladi. U bir nechta maqolalarni nashr etgan va bir nechta xalqaro konferentsiyalarda shaxsiy kompyuter a'zosi va sharhlovchisi sifatida ishlagan jurnallar. Doktor Faiz shuningdek , Karfagen universiteti IHECda “Elektron tijorat va texnologik innovatsiyalar” magistri va “Biznes razvedkasi” magistri uchun mas'uldir .

Kamel Smayli 2002 yildan beri Lotaringiya universiteti professori , 1991 yilda Nensi 1 universitetida nutqni avtomatik aniqlash bo'yicha PhD darajasini oldi . U 2001 yilda HDRni himoya qilgan (Tilni statistik modellashtirish: nutqni tanib olishdan mashina tarjimasigacha). Uning 20 yildan ortiq vaqtdan beri tadqiqotga qiziqishi nutqni avtomatik aniqlash uchun tilni statistik modellashtirish bilan bog'liq. 2000 yildan boshlab u tadqiqotini nutqdan nutqqa tarjima qilishga yo'naltirdi.

U nutqni avtomatik aniqlashga oid bir qancha Yevropa va milliy loyihalarda ishtirok etgan : COCOS, MULTIWORKS, COST, MIAMM, IVOMOB (RNRT loyihasi) va CMCU. U 14 nafar PhD va HDR talabalariga maslahat berdi va Frantsiya, Germaniya, Ispaniya va Jazoirdagi 35 dan ortiq PhD qo'mitalarida qatnashdi .
U bir nechta dastur qo'mitalarida qatnashgan: Interspeech, Eurospeech, ICSLP, ICASSP, TALN, ICWMI, SIIE, TAIMA, Machine Translation, Kompyuter nutqi va tili, Nutq aloqasi, Tabiiy til muhandisligi jurnali,. . .

U Yaponiya, Frantsiya, Tunis, Jazoir va Marokashda taklif etilgan ma'ruzachilar sifatida nutq so'zlash uchun bir necha bor taklif qilingan. U xalqaro konferentsiyalar va jurnallarda 90 ta maqola va frankofon konferentsiyalarida 20 ta maqola chop etgan. Bundan tashqari, janob Smaili 7 yil davomida MIAGE (magistr va bakalavriat) kafedrasi mudiri va 5 yil davomida UFR (fakultetga teng) Matematika va informatika bo'limi boshlig'i bo'lib, u erda 30 dan ortiq doimiy xodimlarni va 120 ta vaqtinchalik lavozimlarni, 500 talabani boshqargan.