Batuhan Duru YELTEKIN
SABIS SUN Beynəlxalq Məktəbi

DODIOM RU: RUS DİLİNDƏ İDİOMALARIN QURULMASI ÜÇÜN OYUNA BƏNZƏR TELEQRAM ÇATBOT

Xülasə

Müasir təbii dil emalının (TDE) tətbiqlərinin (məsələn, Google Translate, Apple Siri və Samsung Bixby) ən çox rast gəlinən çatışmazlıqlarından biri onların çox sözlü ifadələri (ÇSİ) düzgün emal etmək qabiliyyətinin zəif olmasıdır. ÇSİ həm kompozisiya, həm də qeyri-kompozisiya-idiomatik mənaları ilə mövcuddur.

Bu tədqiqatın məqsədi rus dilinin morfoloji təhlilini başa düşmək və idiom korporasiyasının qurulmasına dair ədəbiyyatda ilk dəfə tətbiq olunan və sınaqdan keçirilən kraudsorsinq və izdihamlı reytinq yanaşmasını təqdim etməkdir. Bu məqalə həm də rus dili üçün ən effektiv sonlu vəziyyət çeviricilərindən biri olan UDAR-ı araşdırır. Asan inteqrasiyası üçün bu layihədə Stanza ÇSİ adlı bir alət dəsti istifadə edilmişdir. Stanza ÇSİ çatbot işləyərkən 29 gün ərzində ifadələri effektiv şəkildə lemmatizasiya etməyə kömək etdi. Bu müddət ərzində toplanmış məlumatlar çatbotun effektivliyini nümayiş etdirmək üçün bu hesabata daxil edilmişdir.

Açar sözlər: *Təbii dil emalı, çox sözlü ifadələr, sonlu vəziyyət çeviriciləri, UDAR, oyunlaşdırılmış kraudsorsinq*

UOT: 811; 004,42
DOİ: 10.54414/osex5479

Giriş

ÇSİ-lər "sərbəst söz birləşmələrinə münasibətdə formal və/və ya funksional idiosinkratik xassələri nümayiş etdirərək "vahid" kimi davranan iki və ya daha çox sözdən əmələ gələn dil obyektləri" kimi müəyyən edilmişdir (Masini, 2019). Onlar həm kompozisiya, həm də qeyri-kompozisiya/frasemik ola bilər (məsələn, *ayağı sınmış!* idioması adətən *uğurlar!* mənasını verən idiom kimi ifadə edilir). E-poçt filtrlərindən tutmuş dil tərcümə proqramlarına qədər dəyişən təbii dil emal (TDE) tətbiqləri, ÇSİ-ləri mənalarına görə emal və fərqləndirməkdə bəzi qüsurlara malikdir. Bu çatışmazlıqlar modelləri daha az yetkin olan proqramlarda (yəni, çoxlu məlumatlara öyrədilməmiş) daha çox yayılmışdır, çünki dil modelləri üzərində işlədikləri dilin dərin semantik anlayışı yoxdur (Eryiğit, Şentaş, & Monti, 2022). Yuxarıda qeyd olunan qeyri-adekvatlıqdan qaynaqlanan problemləri həll edən neyron şəbəkələri yaratmaq üçün modellər böyük miqdarda yüksək keyfiyyətli və müxtəlif təlim məlumatlarını tələb edir. Bununla belə, bu təlim məlumat dəstini tapmaq, demək, etməkdən daha asandır. İnternetdə çoxlu cümlələr, o cümlədən dil modellərinin öyrənə biləcəyi ÇSI-lər olsa da, məlumatların keyfiyyəti və müxtəlifliyi ilə bağlı problemlər yaranır. Rus morfologiyasında gözlənilən istisnalar və fonoloji dəyişikliklərlə üç cins, iki cəm və altı hal var (Leaver, Rifkin, & Shekhtman, 2004

UDAR – MÜXTƏLİF DİLLƏRİN FSC-LƏRİ ARASINDA MÜQAYİSƏ

UDAR sonlu vəziyyət analizatoru və onun digər sonlu vəziyyət çeviriciləri ilə müqayisəsi haqqında müzakirəyə davam etməzdən əvvəl mən sonlu vəziyyət çeviricisinin nə olduğu haqqında qısa məlumat verirəm və onun mexanizmini sadə dillə izah edirəm.

Sonlu vəziyyət çeviriciləri (FSCs) bir əsas fərqi ilə sonlu vəziyyət avtomatlarına (FSA) çox oxşardır: FSA-larda yalnız bir yaddaş lenti, FSC-lərdə isə iki **giriş** və **çıxış** lenti var. FSA-lar yalnız girişi oxuyur, FSC-lər isə həm giriş dilini oxuyur, həm də fərqli dildə çıxış yaradır. Bu sonlu vəziyyət maşınlarının istifadəsi buna görə də çox fərqlidir: FSA-lar adətən nümunələri tanımaq üçün istifadə olunur və FSC-lər müxtəlif sətirlər arasında tərcümə etmək üçün istifadə olunur.

Sonlu dəstlərin maraqlı cəhəti (FSC-lər əsasında qurulur) onların real həyat tətbiqlərində gözlənilməz faydalılığıdır. Google Translate və Siri kimi əsas tətbiqləri birlikdə idarə edən kiçik mexanizmlərin əsasını qoyacağını yəqin ki, heç vaxt gözlənməzdi.

UDAR: Rus dilinin sonlu vəziyyət kompleks analizatoru

İndi FSCs haqqında qısa icmalımız tamamlandığından, gəlin tədqiqatımızla daha yaxından əlaqəli olan daha bucaqlı bir şeyi müzakirə edək.

UDAR, ixtiraçısı Robert Reynolds-a görə, vurğulanmış söz formalarını idarə edən rus

dilinin sonlu vəziyyət morfoloji analizatorudur. O, iki səviyyəli formalizmdən lexc və twolc dillərindən istifadə edir. Bu dillər müvafiq olaraq "əsas formaların leksik şəbəkəsini yaratmaq və yaxşı formalaşmış səth formaları yaratmaq üçün əsas formalar üzərində orfoqrafik və morfofonoloji qaydaları həyata keçirmək" üçün istifadə olunur (Reynolds, 2016).

Məsələn, системы (/sʲɪˈsʲtʲemɨ/) sözü üçün UDAR üç leksik analiz yaradır. Birinci təhlil üçün proqram əvvəlcə sözün lemmasını bəyan edir, sonra isə onun isim, cinsi, cansızlığı, çoxluğu və ittiham halı kimi leksik xüsusiyyətlərini bildirir. Digər təhlillər isə системы sözünün müxtəlif xüsusiyyətlərdən ibarət ola biləcəyini iddia edirlər: ikinci təhlil sözün nominal halda ifadə edildiyini iddia edir, üçüncü təhlil isə sözün tək isim olduğunu və cinsiyyət halında ifadə olunduğunu bildirərək daha da ziddiyyətli bir etiraz edir.

Son iki onillikdə bir neçə rus morfoloji analizatorunun inkişafı müşahidə edildi. Bu analizatorların çoxu öz yolları ilə faydalıdır; lakin, onlar hələ də onları güclü alətlər edəcək bəzi atributlara malik deyillər. UDAR isə bu vasitələrin ən cəlbedici komponentlərini əhatə edir və onlara bir paketdə xidmət göstərir. Aşağıdakı cədvəl UDAR-ın digər analizatorlara nisbətən üstünlüklərini göstərir.

Stanza alət dəsti və lemmatizasiya prosesi

UDAR kimi sonlu vəziyyət çeviriciləri və analizatorları sözlərin leksik xüsusiyyətlərini müəyyən etmək və sənədləşdirmək üçün faydalıdır. Rus dilinin analizatorları xüsusilə təsir edicidir, çünki onlar morfoloji cəhətdən mürəkkəb rus cümlələrindən sözləri müəyyən xüsusiyyətlərə bölmək kimi çətin bir iş görürlər. Bu alətlər və ya ən azı onların arxasında duran nəzəriyyə, çatbot üçün təqdimetmənin doğrulama sisteminin yaradılmasında istifadə olunan Stanza kimi son dərəcə güclü NLP alət dəstlərinin əsasını qoyur

Stanza "bir çox insan dillərinin linqvistik təhlili üçün dəqiq və səmərəli alətlər toplusudur" (Qi və digərləri, 2020). Bu, əsasən, cümlələri və sözləri, onların lemmalarını (əsas formalarını), nitq hissələrini və morfoloji xüsusiyyətlərini müəyyən edə bilən alətlər dəstidir.

Lemmatizasiya prosesinin texniki məlumatı

id		4
date		2022-05-21
name		белая ворона
meaning		уникальный персонаж
language		ENGLISH
lemmas		{белый,ворона}
words		{белая,ворона}
category		VID
video_link		https://www.youtube.com

Save | Save and continue edit | Delete

: Nümunə MWE verilənlər bazasındakı ID nömrəsi, təyin olunduğu tarix, rus dilində idiomun mənası, lemmaların siyahısı və MWE-dəki sözlərin siyahısı, habelə və MWE-nin idiomatik mənasını izah edən videoların keçidi.

Bütün MWE-lər, təqdimetmə, nəzərdən keçirmə və istifadəçi məlumatları təhlükəsiz PostgreSQL verilənlər bazasında saxlanıldı. Bu verilənlər bazası vasitəsilə biz müəyyən bir tarixə MWE təyin edə bildik ki, bot həmin MWE-ni həmin tarixdə oyunçulara göndərə bilsin. MWE ilə yanaşı, bəzi əlavə məlumatlar (istifadəçilərə də göndərildi), daha dəqiq desək MWE-nin idiomatik mənası və onu izah edən YouTube videosu saxlanıldı. Bununla belə, istifadəçilərlə paylaşılmayan bir məlumat Şəkil 3-də göstərildiyi kimi MWE-nin lemmalarının siyahısı (*dil* sətiri nəzərə alınmır) da var idi.

DODIOM RU çatbot dizayni və geympley

Dodiom çatbotu çatbot üzrə araşdırmanın ilk buraxılışında türk, ingilis və italyan dillərində lokallaşdırılıb. Bununla belə, rus dilinin xüsusi olaraq fərqli kiril qrafikasından istifadə etdiyi və daha mürəkkəb morfologiyaya malik olduğu üçün rus dilinin lokallaşdırılmasına dair ayrıca tədqiqat işi lazım bildi.

Əslində, oyun oyunçuları cümlələr yazmağa sövq edir, o cümlədən oyunun işlədiyi hər gün

oyunçuya göndərilən MWE. Oyunçu, öz növbəsində, günün MWE-si daxil olmaqla cümlələr təqdim edir və onun hərfi və ya idiomatik mənada ifadə edilib-edilmədiyini təsnif edir. Oyunçular digər oyunçuların təqdimatlarını nəzərdən keçirə və müsbət və ya mənfi reytinqlər buraxa bilərlər. Oyunçu daha çox nümunəni nəzərdən keçirdikcə, daha çox xal qazanır və təqdimat müəllifi daha çox müsbət rəylər qazandıqca, onlar da daha çox xal qazanırlar. Adətən, rəyçi verdiyi hər rəyə görə bir xal, müəllif isə təqdimatlarından birində aldığı hər müsbət rəyə görə on xal alır. Çatbotun iş saatlarının sonunda (həftənin bütün günləri 11:00-dan 23:00-a qədər) ən çox xal toplayan istifadəçi günün çempionu sayılır və pul dəyəri ilə mükafat alır Gün ərzində müəyyən sayda təqdimat göndərmək kimi hədləri keçən oyunçular nailiyyətlər qazanırlar. Çatbot Azərbaycanda istifadəyə verilib və gündəlik mükafat Wolt qida çatdırılması proqramından 15₼ hədiyyə kartı olub.

§Yuxarıdakı oyunun izahı kraudsorsinqin tərifinə uyğun gəlir. Belə ki, bu, ödəniş müqabilində layihəyə töhfə vermək üçün "səpələnmiş" insanların böyük bir qrupunun işə götürülməsi deməkdir. O, həmçinin rəqabətli oyunun strukturuna uyğundur; buna görə də ona oyunlaşdırılmış çatbot deyilir.

Qeyd etmək vacibdir ki, hər bir yeni oyunçu botun tədqiqat məqsədləri üçün hazırlandığı və təqdim etdikləri nümunələrin təlim modelləri üçün istifadə oluna biləcəyi barədə məlumatlandırılmışdır. Onlara heç bir şəxsi məlumatın işlənilmədiyi və ya ictimaiyyətə açıqlanmadığı da bildirilib.

Əldə olunmuş məlumatlarin təhlil edilməsi

Oyunun fəaliyyəti dayandırıldıqdan sonra bütün müvafiq məlumatlar Python-un Matplotlib kitabxanasından istifadə edilərək çıxarıldı və vizuallaşdırıldı. Bu vizualizasiyalarda Dodiom RU-dan alınan məlumatlar çatbotun ingilis dilindəki versiyasından (həmin ilin fevralında fəaliyyət göstərib) verilənlərlə müqayisə edilib. Qeyd etmək vacibdir ki, Dodiom RU-da bütün gündəlik oyunların çempionları pul dəyəri olan əşyalarla mükafatlandırılsa da, Dodiom EN-nin fəaliyyət müddəti iki mərhələyə bölünmüşdü və birinci mərhələdə heç bir pul dəyərində mükafat verilməmişdir. Bu kontekstdə Dodiom EN-nin birinci mərhələsi "EN", ikinci mərhələ isə "ENw/Mon.Reward" olaraq etiketlənir.

Nəticə

Yuxarda qeyd olunduğu kimi, bütün dünyada dillər qeyri-müəyyənlik atributunu inkişaf etdirir. Qeyri-müəyyənlik dil modellərinin MWE-ləri emal etməkdə çətinlik çəkməsinin səbəbidir, burada ifadələr bəzən qeyri-kompozitivdir. Bu modellər üçün daha yaxşı təlim məlumatları əldə etməklə bu problemi yaxşılaşdırmaq olar. Oyunlaşdırılmış kraudsorsinq "kütlənin müdrikliyindən" istifadə etməklə və yüksək sayda keyfiyyətli nümunələr əldə etmək üçün rəqabəti təşviq etməklə bu ehtiyacı ödəyə bilər. İstifadəçi və təqdimetmə/nəzarət statistikası bu fikri nümayiş etdirməyə xidmət edir. Verdikləri töhfələrə görə ən yaxşı oyunçular hədiyyə kartları ilə həvəsləndirildi və bu, böyük motivasiya mənbəyi oldu.

Bu tədqiqat işi Dodiom çatbotunun ixtiraçılarının tədqiqat işinin yekununda müəyyən etdiyi məqsədə qismən nail olur: oyunun əhatə dairəsini digər dillərə genişləndirmək. Ümid edirik ki, bu oyun gələcək illərdə daha çox dildə mövcud olacaq və bu tədqiqat sahəsinin inkişafına töhfə verəcək.

ƏDƏBİYYAT SİYAHISI

Chomsky, N., Rizzi, L., & Belletti, A. (2002). An interview on minimalism. In N. Chomsky, *On Nature and Language* (pp. 92-161). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511613876

Eryiğit, G., Şentaş, A., & Monti, J. (2022). Gamified crowdsourcing for idiom corpora construction. *Natural Language Engineering*, 1-–33.

Leaver, B. L., Rifkin, B., & Shekhtman, B. (2004). Apples and oranges are both fruit, but they do not taste the same: A response to Wynne Wong and Bill VanPatten. *Foreign Language Annals*, 125–132.

Masini, F. (2019, September 30). Multi-word expressions and morphology. In *Oxford research encyclopedias*. https://oxfordre.com/linguistics/view/10.1093/acrefore/9780199384655.001.0001/acrefore-9780199384655-e-611?print=pdf

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python

natural language processing toolkit for many human languages.

Reynolds, R. J. (2016). Russian natural language processing for computer-assisted language learning. Tromsø: UiT The Arctic University of Norway.

R, N., Parimi, M. R., S, A. R., NS, N. K., & Babu, S. (2020). Develop CSR themes using text-mining and topic modelling techniques. In *I. Staff, 2020 IEEE International Conference on Cloud Computing in Emerging Markets* (CCEM) (pp. 67-71). Bengaluru: IEEE.

Батухан Дуру Елтекин
Международная школа САБИС САН

DODIOM RU: ИГРОВОЙ ТЕЛЕГРАМНЫЙ ЧАТ-БОТ ДЛЯ ПОСТРОЕНИЯ ИДИОМ НА РУССКОМ ЯЗЫКЕ

Резюме: Одним из наиболее распространенных недостатков современных приложений обработки естественного языка (ОЕЯ) (таких как Google Translate, Siri от Apple и Bixby от Samsung) является их неразвитая способность правильно обрабатывать многословные выражения (МСВ). МСВ существует как в композиционном, так и в некомпозиционно-идиоматическом значениях.

Целью данного исследования является понимание морфологического анализа русского языка и представление краудсорсинга и краудрейтингового подхода, впервые примененного и апробированного в литературе по построению корпусов идиом. В этой статье также рассматривается УДАР- один из наиболее эффективных преобразователей с конечным числом состояний для русского языка.

Для упрощения интеграции в этом проекте использовался инструментарий под названием Stanza MCB. Stanza MCB помогла эффективно лемматизировать фразы за 29 дней, пока работал чат-бот. Данные, собранные за этот период, включены в этот отчет, чтобы продемонстрировать эффективность чат-бота.

Ключевые слова: обработка естественного языка, многословные выражения, преобразователи конечной ситуации (ПКС), UDAR, геймифицированный краудсорсинг.

Batuhan Duru Yeltekin
International School of SABIS SUN

DODIOM RU: A GAME-LIKE TELEGRAM CHATBOT FOR BUILDING IDIOMS IN THE RUSSIAN LANGUAGE

Summary: One of the most common shortcomings of modern natural language processing (NLP) applications (such as Google Translate, Apple's Siri, and Samsung's Bixby) is their poor ability to properly process multi-word expressions (MWEs). MWEs exists with both compositional and non-compositional-idiomatic meanings. The aim of this study is to understand the morphological analysis of the Russian language and to present a crowdsourcing and crowd-rating approach that is first applied and tested in the literature on idiom corpora construction. This paper also investigates UDAR - one of the most efficient finite-state converters for the Russian language.

For an easy integration, a toolkit called Stanza MWEs was used in this project. Stanza MWEs helped to effectively lemmatize phrases in 29 days while the chatbot was running. The data collected during this period is included in this report to demonstrate the effectiveness of the chatbot.

Keywords: Keywords: *Natural language processing, multi-word expressions, final situation converters, UDAR, gamified crowdsourcing*

Daxil olub: 11.08.2022