Петя Осенова[1]
(СУ „Св. Кл. Охридски", България)

ТУИТОВЕТЕ КАТО ПРЕДИЗВИКАТЕЛСТВО ПРЕД АВТОМАТИЧНАТА ЛИНГВИСТИЧНА ОБРАБОТКА

Petya Osenova

TWEETS AS A CHALLENGE FOR THE AUTOMATIC LINGUISTIC PROCESSING

Резюме: Обект на статията са особеностите на писмената разговорна реч в туитовете като предизвикателство за автоматичния езиков анализ. Този анализ включва: сегментиране на текста на думи; морфологичен анализ по части на речта и техните граматични характеристики; дependентен синтактичен анализ; разпознаване на имена на хора, локации, организации; разпознаване на абревиатури. Проблемите, които възникват, са свързани предимно: с думи, които не са в речника; със сливане на думи; с разговорни варианти на изрази без нормализация към книжовни съответствия и др. Изследването използва 630 туита на български език, свързани с банковата криза на КТБ и ПИБ през 2014 г.

Abstract: The paper focuses on the specificities of the written colloquial speech in tweets as a challenge for the automatic linguistic analysis. Such an analysis includes: text segmentation into words; morphological analysis in parts-of-speech and related grammatical characteristics; dependency syntactic analysis; named entity recognition of people, locations and organizations; handling abbreviations. The problems are of the following kinds: out-of-vocabulary words; word blending; colloquial variants that have not been normalized, etc. The survey explores 630 tweets that discuss the crisis of two banks in Bulgaria in 2014.

Ключови думи: туит, български език, автоматична лингвистична обработка, писмена разговорна реч

[1] petyaosenova@slav.uni-sofia.bg

Key words: tweet, Bulgarian language, automatic linguistic processing; written colloquial speech

Туитът е кратка форма на комуникация с ограничен брой символи в социалната мрежа Туитър. Това означава, че той е микропослание, което предава най-важната информация, за разлика от възможността за писане на по-големи текстове във Фейсбук например. Тази кратка форма обаче придобива силата на свързания текст чрез последователностите от коментари по определена тема (т. нар. нишки, threads). Туитът има двойна функция: от една страна, чрез него може да се изрази официална позиция на дадено лице. Тогава се активират особеностите на писмената реч. От друга страна, чрез него може да се изрази и по-неформална позиция. Тогава се активират особеностите на устната реч. В тази статия се разглежда предимно вторият тип изразяване, при който се използва разговорна форма на общуване с всички нейни специфики на контрахираност, недовършеност, повторителност и т.н. Друга особеност на туита е, че той разчита изключително много на хаштага (#) за тематичната категоризация на микротекстовете. Самите туитове са голямо предизвикателство пред автоматичната обработка на текст, но съдържанието на хаштаговете също е проблем, защото то представлява изкуствени дълги думи, получени от речниковите думи. В този текст обаче не се разискват проблемите на автоматична обработка в рамките на хаштаговете. Това е отделна и никак нелесна задача.

Жизненият цикъл на туита може да се обобщи по следния начин: а) написването и публикуването на съобщението в Мрежата; б) разпространяването му (чрез т. нар. повторено копиране, re-tweet) и накрая в) получаването на различни реакции като отговор на написаната позиция.

В изследването са използвани 630 туита (или 19 269 думи) на български език, свързани с темата за банковата криза с КТБ и ПИБ през 2014 г. Това е извадка от по-голям корпус от около 16 000 туита, който беше събран по европейския проект Pheme[2] (2014–2017). Проектът имаше за цел да създаде технология за бърз и ефективен анализ на големи масиви от съдържание, генерирано от различни потребители в социалните мрежи с оглед на показатели като истинност на съдържанието и доверие към него. На заден план останаха

[2] https://www.pheme.eu/

фактори като големината на данните; бързината на разпространение на туитовете и тяхното разнообразие. За да се стигне до анализ на съдържанието обаче, беше нужно да се обработят автоматично туитове на различни езици, сред които и на български.

Туитовете бяха обработени с автоматичния процесор за български език (Savkov et al. 2012). Този процесор включва следните компоненти: сегментатор; морфосинтактичен анализатор; модул за снемане на многозначността; лематизатор; модул за разпознаване на имена и абревиатури, както и синтактичен анализатор. Сегментаторът разделя текста на думи, пунктуационни знаци и символи. За писмен[3] текст той работи с точност от около 90%. Морфосинтактичният анализатор определя частите на речта на думите, както и набора от техните граматични характеристики (ако ги има). За писмен текст той постига резултат от около 97%. Модулът за снемане на многозначност в случая е част от морфосинтактичния анализатор. Лематизаторът привежда дадената словоформа в основната лексема. За писмен текст той има точност от около 95%. Успехът на тази стъпка зависи от точността на морфосинтактичния анализатор заедно с модула за снемане на многозначност. Анализаторът разчита и на информацията, кодирана в голям морфологичен речник на българския език. Модулът за разпознаване на имена и абревиатури може да бъде част от сегментатора, но може да бъде и самостоятелна програма за анализ. Този модул работи с точност от около 80% върху писмени текстове. Дependентният синтактичен анализ постига резултат от около 93% върху писмен текст.

Когато процесорът извърши автоматичния анализ на корпуса от туитове, впечатление направиха следните проблеми: при сегментацията се получи около 2% грешка; при новите думи, които не се срещат в морфологичния речник, грешката беше от около 1%. Морфосинтактичният анализатор направи грешка от около 2,26%. Оттам грешката се пренесе при лематизатора и съответно – при синтактичния анализатор, като там тя нараства до 4–5%. Както се вижда, това не са големи проценти грешка, но трябва да се има предвид, че и извадката от туитове не е много голяма. При обработка на големи данни процентът на грешките може да нарасне още.

[3] Тематичните области, върху които са оценени модулите, са основно медийни текстове и художествена литература.

Ясно е, че основните проблеми при автоматичния анализ на туитове се дължат на особеностите на тази форма на комуникация. Това са: наличието на думи от чужд език; използването на много емотикони вместо пунктуация; недостатъчният контекст на употреба на дадена дума, поради което програмата взема неправилно решение; невъзможността да се отдели хаштагът от изказването и др. Извън тези особености туитовете споделят общите проблеми при анализ на разговорна реч. Те са формулирани като маркери на разговорната реч в Радева (2012). От тези маркери най-съществени за дискутираната жанрова форма на туита са следните: преобладаване на експресивната лексика и прагматичните частици; недовършеност/елипси/кондензация/разкъсване на структурите; повторения; чужди думи, предадени на български език. В по-малка степен се наблюдават маркери като: преобладаване на сложни части (*бие я парата = има много пари*); дублирани конструкции; клишета; умалителни/огрубителни думи.

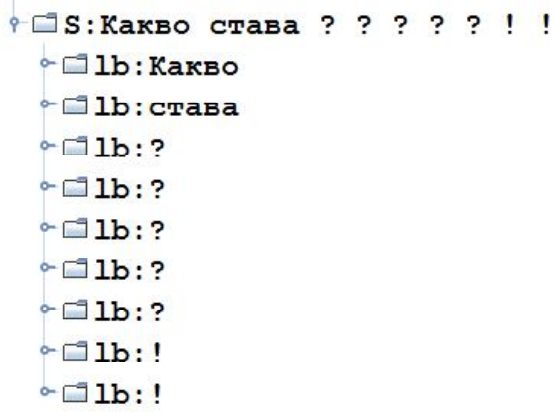
Особеностите на устната разговорна реч от фонетична и морфологична гледна точка са описани подробно в Джонова, Велкова (2014), а от синтактична гледна точка – в Тишева (2014а и 2014б). От фонетична гледна точка и при туитовете се наблюдава засилена степен на елизии, а от морфологична – липсата на падеж при въпросителните местоимения и производните от тях местоимения. От синтактична гледна точка припокриване с особеностите на устната реч има най-вече по отношение на дислоцирания словоред и прагматичните маркери.

Нека разгледаме поотделно проблемите при всяка стъпка от автоматичния анализ на процесора, като комбинираме стъпката с наличието на съответните разговорни маркери. Обработеният текст е представен в системата CLaRK[4].

Сегментиране на текста. При тази начална стъпка проблем представляват: смесването на кирилски и латински букви; свободното използване на пунктуация или липсата на такава; наличието на емотикони (включително и вместо пунктуация); наличието на линкове; изкуствените думи и др. На фиг. 1. е представено изречение, което

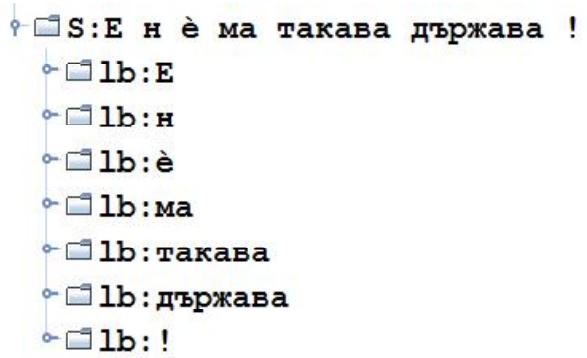
[4] http://bultreebank.org/en/clark/

завършва с четири въпросителни и два удивителни знака. Макар че сегментаторът ги е разделил коректно, на следващите равнища на анализ би било по-добре, ако те са разгледани заедно. Тук обаче е трудно предвидимо колко и какви знаци ще използва потребителят.



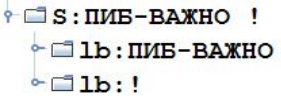
Фиг. 1. Пунктуация

На фиг. 2. е представено изречение, при което е използван символ, който не е кирилски. Това е попречило на сегментатора да разпознае думата *нема* като едно цяло. Друг е въпросът, че на следващото равнище на анализ сигурно думата не би била разпозната като глагол без нормализация към нормативната форма *няма*.



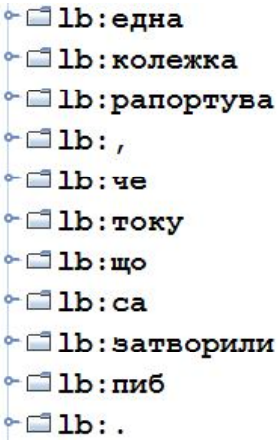
Фиг. 2. Смесване на символи

На фиг. 3. е представена изкуствената сложна дума *ПИБ-ВАЖНО* поради грешно изписан дефис вместо голямо тире:



Фиг. 3. Изкуствени сложни думи

На фиг. 4 е представено сгрешено сегментиране поради грешно изписване на сложна дума като две думи – *току* и *що* – вместо полуслятото *току-що*:



Фиг. 4. Изкуствени несложни думи

Морфосинтактичен анализ. Основен проблем при този тип анализ не са думите и словоформите, които липсват в речника. При подобни случаи с добра точност се приписва съответната част на речта заедно с граматичните ѝ характеристики. По-проблемни за правилното определяне на частта на речта са: съкращенията (*кардио* вм. *кардиологичен*), слетите думи (*фзатвора* вм. *в затвора*; *изкефиме* вм. *изкефи ме*), грешките (*триасе-четирасе* вм. *трийсет-четирийсет*; *бес* вм. *без*); разговорните форми (*тея* вм. *тези*; *наайс* вм. *хубаво*), сред които особено удължаванията на гласни (*лелее* вм. *леле*); функционалните думи като междуметия и частици. Трудностите са

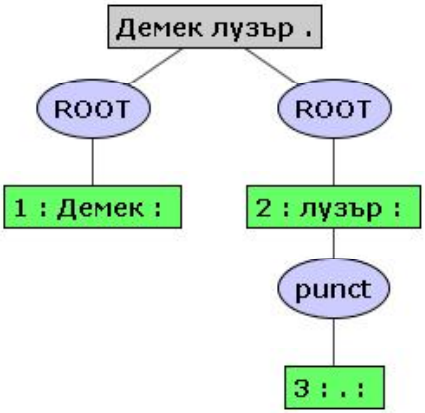
в две посоки: едната е грешен или нетипичен низ от букви в думата, а другата е създаване на допълнителни многозначности. В рамките на първата посока във въпроса *Как сиииииии?* глаголът за 2 л., ед. ч., сег. вр. *сиииииии* е разпознат като съществително вместо като глагол; в израза *Как се превеждаше тва?* контрахираното показателно местоимение *тва* по някаква причина е разпознато като глагол в 3 л., ед. ч., сег. вр. По отношение на втората посока може да се даде следният пример: в конструкцията *Лелее, ужас* думата *лелее* е разпозната не като междуметие, а като форма за 3 л., ед. ч. на глагола *лелея*.

Лематизация. Това е процедура, при която словоформата се привежда в своята основна форма. Например словоформата *дойдоха* се лематизира в *дойда*. Тази стъпка зависи изключително много от правилното определяне на част на речта. Например частицата *бре* е разпозната грешно като глагол. Затова и като лема е копирана пак тя, тъй като не е намерена основна форма на подобен глагол. При грешни форми на дадена дума също не се стига до основната ѝ форма. Например в израза *са СПРЯЛИ всякакви плащания* формата *спряли* не се разпознава като форма на глагола *спра*. В изречения като *Как се превеждаше тва, а да да, ЗАГУБЕНЯК* първо е сгрешена думата *тва*. Тя е разпозната като глагол, а не местоимение, както беше споменато и по-горе. След това при двете *да* първото грешно е разпознато като глаголна частица, а второто правилно – като утвърдителна частица. В израза *офсете* [вм. овцете] *си върнат парите* думата *офсете* е разпозната като звателна форма на съществителното от м. р. *офсет*. Разговорното *сори* в смисъл на *съжалявам* е разпознато като фамилно име.

Модул за разпознаване на имена и абревиатури. При тази стъпка се разчита и на готови списъци с имена и абревиатури, и на правила за определени графични характеристики (акронимите най-често се пишат с главни букви; името започва с главна буква и т.н.). Тук, разбира се, също има доста проблеми. Например графичните съкращения могат да съвпадат с думи (срв. *ген* и *ген.=генерал*); думите могат да съвпадат с имена (срв. името *Сам* и прилагателното *сам*) и др. Неразпозната остава например фразата, при която няма точка след графичното съкращение *На бул България*. Първите две думи са разпознати грешно като едно изречение – *На бул*, а *България* – като второ изречение.

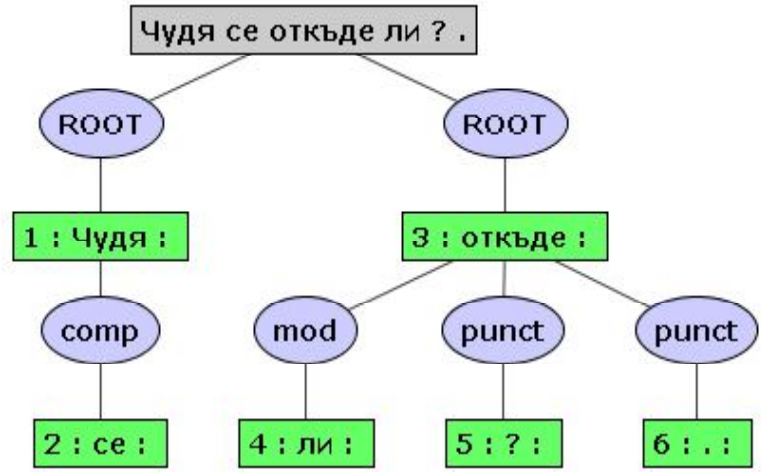
Синтактичен анализ. Този тип анализ изключително много зависи от предходните стъпки и най-вече от правилния морфологичен анализ, защото без частите на речта не могат да се осигурят правилните връзки между опора и зависима част. В случая е направен дependентен автоматичен анализ, т.е. анализ, при който определящи са зависимостите между думите, а не фразите.

Освен влиянието на предхождащите анализи обаче значение имат и други фактори. Това са: голямото количество безглаголни изречения; недовършените изречения; неправилно разделените изречения; повторенията и др. Например в изречението *Демек лузър*, макар че двете думи са определени правилно на морфологично равнище, синтактичният анализ е грешен. Той разглежда и двете думи едновременно като корени (ROOT) на изречението. Но изречението може да има само един корен. Срв. фиг. 5.:



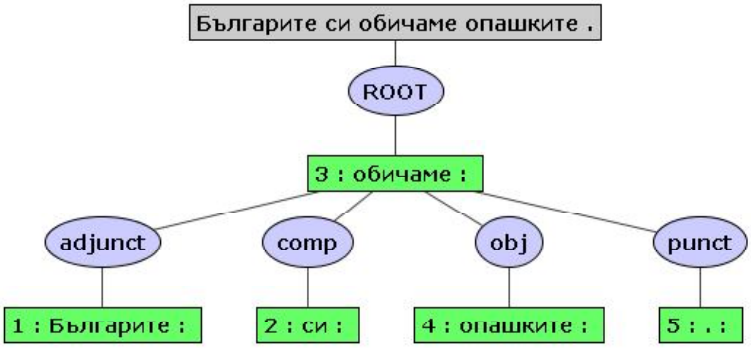
Фиг. 5. Изречение с два корена

Подобен проблем има и в изречение с елипсирани елементи: *Чудя се откъде ли* [*народът се е сдобил с пари*]? Автоматичният анализатор отново поставя два корена. Срв. фиг. 6.:



Фиг. 6. Сложно съставно изречение с елипсирани елементи в подчиненото изречение

Грешен анализ може да се получи и при особени случаи на съгласуване. Срв. фиг. 7.:



Фиг. 7. Изречение с особено съгласуване между подлог и сказуемо

В този случай съществителното *Българите* не е анализиарно като подлог, а като адюнкт най-вероятно поради факта, че липсва местоимението *ние*, което реално се съгласува със сказуемото.

Дотук бяха набелязани някои основни проблеми при автоматичната обработка на туитове на български език.

Интересно е да се види обаче и честотата на срещане на словоформите, представена в Таблица 1.:

Таблица 1. Честота на срещане на словоформите

СЛОВОФОРМА	ЧЕСТОТА
пари	66
пред	65
днес	57
парите	56
ли	54
КТБ	53
ми	49
клиенти	47
няма	45
не	42
банка	41
има	40
☺	37
млн	36
лв	34
банката, слухове, това	33
понеделник	31
време, опашка, работи	29

Прави впечатление, че освен съществителните имена за пари, КТБ, клиенти, банка, се използват и маркери за пространство и време – локативният предлог *пред,* съществителните *опашка* и *понеделник*. От честотата на въпросителната частица *ли* става ясно, че

голяма част от туитовете са всъщност въпроси. Има висока употреба на отрицателни форми и частици, както и на емотикони. По отношение на местоименията в разгледаните туитове е интересно, че с висока честота са само два вида: дателното местоимение в 1 л., ед.ч. *ми*, както и показателното местоимение за ср.р., ед.ч. *това*. Според изследването на Илиева (2005: 66) в книжовно-разговорната реч личното дателно местоимение *ми* е четвърто по честота на употреба. Сред често употребяваните се оказва и местоимението *това* – на дванайсето място. Липсата на местоимението *аз*, което при Илиева води класацията по честота, би могло да се обясни с факта, че от една страна, темата за кризата с банките не предполага *аз*-центричност, и от друга, наличието на нулева субектност е много по-висока в писмената разговорна реч в сравнение с устната. Липсата на третоличните местоимения, които при Илиева са на второ и трето място по честота (*то* и *той*), може да се обясни с употребата на имената на банките вместо анафоризиращите елементи.

В заключение може да се обобщи, че проблемите, които поставят туитовете пред автоматичната лингвистична обработка, споделят общите черти на проблемите при обработка на корпуси с разговорна реч (била тя парламентарна, от блогове, в диалогов режим и под.), но те носят и своите специфики. Тези специфики се отнасят най-вече до краткостта на жанра; системната употреба на хаштагове, които прекъсват съдържанието; смесването на символи и азбуки; наличието на изразен фокус върху определена тема за дискусия; динамичността при обмяната на информация.

ЛИТЕРАТУРА

Джонова, Велкова 2014: Джонова, М., Й. Велкова. Фонетични и морфологични особености на устната реч. // *Как говори съвременният българин: Граматика и устна реч*, том 1, ФНИ и фондация Фокус, 104–156. **Dzhonova, Velkova 2014:** Fonetichni i morfologichni osobenosti na ustnata rech. // Kak govori savremenniyat balgarin, tom 1, FNI i fondatsia Fokus, 104–156.

Илиева 2005: Илиева, М. *Количествен анализ на местоименната употреба в стиловете на българския език*. Варна: Славена. **Ilieva 2005:**

Kolichestven analiz na mestoimennata upotreba v stilovete na balgarskiya ezik. Varna: Slavena.

Радева 2012: Радева, П. *Динамика в синтаксиса на съвременния български език*. В. Търново: УИ „Св. св. Кирил и Методий", **Radeva 2012:** Radeva, P. Dinamika v sintaksisa na savremenniya balgarski ezik. V. Tarnovo: UI "Sv. sv. Kiril i Metodiy".

Тишева 2014а: Й. Тишева. Синтактични особености на устната реч. // *Как говори съвременният българин: Граматика и устна реч*, том 1, ФНИ и фондация Фокус, 157–180. **Tisheva 2014a:** Sintaktichni osobenosti na ustnata rech. Kak govori savremenniyat balgarin, tom 1, FNI i fondatsiya Fokus, 157–180.

Тишева 2014б: Тишева, Й. Прагматика и устна реч. // *Как говори съвременният българин*, том 2, ФНИ и фондация Фокус. **Tisheva 2014b:** Pragmatika i ustna rech. // Kak govori savremenniyat balgarin, tom 2, FNI i fondatsiya Fokus.

Savkov et al. 2012: Savkov, A., L. Laskova, S. Kancheva, P. Osenova and K. Simov. Linguistic Analysis Processing Line for Bulgarian. // N. Calzolari, K. Choukri, T. Declerck, M. Upur Dopan, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis (eds.) *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey.