МЕТОДИКА ЗА АВТОМАТИЗИРАНО ИЗГРАЖДАНЕ НА ЧЕСТОТЕН РЕЧНИК НА НОВОГРЪЦКИЯ ЕЗИК

Румен Рикевски

1. Увод

Честотният речник е списък от думи, където всяка дума има индикатор за това колко пъти е срещната в даден текст.

Един важен въпрос за езиковите изследователи и лингвистите е до каква степен може да се преведе даден текст с помощта на определен брой лексеми. В изследвания, свързани с английския, немския и испанския езици е установено, че най-употребяваните 1000 лексеми покриват 72–89% от всички думи в текста, като девиацията зависи от вида и жанра на текста. Следващият блок от 1000 лексеми покриват 4–10% от останалите думи, а за третия блок от 1000 лексеми този процент пада на 2–4%. Тези данни, от една страна показват колко важен за даден език е правилно съставения честотен речник, а от друга – дават възможност да се прецени колко време и усилия трябва да се изразходват за научаването на определен корпус от основни за езика думи и колко – за по-малко употребяваните. На базата на честотния речник, след това могат да се съставят най-подходящите учебници и пособия, които оптимално да подпомогнат обучаемите при усвояване на чуждия език. Автоматизираното изграждане на честотен речник става с помощта на една интердисциплинарна теоретико-приложна наука – компютърната лингвистика.

2. Теория

Компютърната лингвистика се занимава както с формалното описание на естествения език, така и с разработването и прилагането на компютърните технологии при статистическото и логическото му анализиране и моделиране. С помощта на компютърната лингвистика се разработват езикови приложения и системи за работа с текстове, като:

– програми за автоматично коригиране на правописа;

– програми за автоматичен превод от един език на друг;

– програми за категоризиране и резюмиране на документи;

– програми за преобразуване на текст в реч и обратно;

– програми, обслужващи лингвистичните изследвания и анализи и др.

За тези цели се появява нужда от по-развити, често пъти изследователски методи и алгоритми за работа с текстове, които максимално да наподобяват разбирането на езика от човека, т.е. появяват се методите на Изкуствения Интелект. Един от тези методи е автоматичният анализ. Той може да обхваща различни нива от заложената в текста информация, а именно – лексикално, граматично, морфологично, синтактично, семантично, контекстно. Това определя множество подзадачи и модули като:

– токънизиране (разделяне текста на определени единици – фонеми, морфеми, графични думи, лексеми, изречения и др.);

– тагиране (приписване на характеристики на всяка отделна единица – приписване на дадени морфологични, синтактични, морфосинтактични, семантични и др. характеристики);

– парсиране – морфологичен анализ, синтактичен анализ, разрешаване на различни езиково специфични явления като анафори, елипси и др. и на различните типове езикова многозначност.

3. Морфологично описание на новогръцкия език

В новогръцкият език има 11 основни категории от думи, които се наричат части на речта.

Пет от частите на речта са неизменяеми: наречие, предлог, съюз, междуметие и частица, а останалите 6 са изменяеми: определителен член, съществително, прилагателно, глагол, местоимение и причастие.

Неизменяемите думи са тези, които не могат да менят формата си, т.е. имат само една единствена форма, например: τώρα, από, δηλαδή, ας, αχ.

Изменяемите части притежават определен брой граматически категории. За определителния член, съществителното, прилагателното, местоимението и причастието това са най-често род, число, падеж; за глагола – лице, число, залог, време, наклонение и др. Изменяемите части на речта менят формата си посредством формообразуващи морфеми (окончание, основа, аугмент) и по такъв начин изразяват различни граматически значения. Така всяка изменяема дума притежава различен брой словоформи, например: τρέχω, τρέχεις, τρέχει, έτρεξα, τρέξαμε...; νέος, νέα, νέο...; όνομα, ονόματος, ονόματα...; αυτός, αυτή, αυτό.... Една от

словоформите се приема за основна (лема). В почти всички речници се дава само тази основна форма. Обикновено това е най-простата словоформа. В новогръцкия език за основна словоформа се избира тази в единствено число, именителен падеж при съществителните и словоформата в изявително наклонение, деятелен залог, сегашно време, първо лице, единствено число при глаголите.

Броят на словоформите за различните части на речта е пряко свързан с граматическите им категории. Словоформите се отличават една от друга поне по едно граматическо значение, т.е. те представляват различни граматически категории, въпреки че имат едно и също лексикално значение. Най-многобройни форми има глаголът, защото той притежава най-много граматически категории.

4. Морфологичен анализатор и лексикална база от данни

Основната трудност в нашия проект се състои в намирането на лемата на постъпилата от даден текст словоформа. Този проблем се разрешава с помощта на морфологичен анализатор, който сравнява и анализира словоформата спрямо една предварително създадена лексикална база от данни.

Лексикалната база данни представлява множество от лексикографско подредени, тагирани (с приписани, езикови характеристики) думи. Изграждането ѝ представлява сложен и трудоемък процес, който трябва да разреши 2 основни проблема: намиране на формални модели, позволяващи думите да бъдат разделяни във формални езикови класове и автоматичното причисляване на дадена дума към един, или друг езиков клас. Формализираните речници се изграждат на базата на формални характеристики на думите и информацията е във вид на формални структури, удобни за компютърна обработка.

Всяка дума която трябва да бъде разпозната, най-напред се разделя на съставните си морфеми и след това се търси в речника. Формализираният речник съдържа морфеми (основа, окончание и аугмент) и една система от правила (формални характеристики), която да описва морфологичните промени на склоняемите части от речта . Примерно за думата „δάσκαλος" регистрираме 8-те окончания в 4-те падежа в ед. и мн.ч. (-ος, -ου, -ο, -ε, -οι, -ων, -ους, -οι) и ги обозначаваме като правило Х, т.е. с помощта на това правило се образуват останалите словоформи на „δάσκαλος". В речника ще впишем основата „δάσκαλ", която представ-

лява шаблон за лексемата „δάσκαλος" + правилото X. Ако сега трябва да впишем думата „άνεμος", която има същото склонение, ще отбележим просто „άνεμ" + X. Същевременно за думата „καφές" ще отбележим „καφ" + Y, където Y е правило означаващо 8-те окончания (-ες, -ε, -ε, -ε, -εδες, -εδων, -εδες, -εδες), а за „καναπές", която се скланя аналогично на „καφές" - „καναπ" + Y. По този икономичен начин вписваме една лексема чрез нейната основна форма (лема), като същевременно чрез системата от формални характеристики сме закодирали всичките ѝ възможни словоформи.

Защо се налага такава една икономичност при съставянето на лексикалната база данни? Знаем, че при компютърните системи въпросите, свързани с памет и бързодействие винаги са от първостепенна важност. Ние, разбира се, можем да използваме друг тип речник, който да съдържа всички възможни словоформи, напр. (δάσκαλος, δασκάλου, δάσκαλο, δάσκαλε...). Обаче, едно е да съставиш и реализираш речник с 50 0000–1 000 000 думи и съвсем друго речник с 50 000–100 000 думи и система за описание на морфологичните промени, състояща се от примерно 1000 правила. По-същия начин има голяма разлика дали търсиш една дума в база данни от 1 000 000 или 100 000 думи. А при прилагателните имена и особено при глаголите в гръцкия език словоформите са доста повече, така че съотношението от думи във формализирания и словоформения речник може да стане още по-голямо.

Строежът на формализирания речник трябва да позволява бърз достъп до търсената дума и лесно добавяне на нови думи. За този случай най-подходяща би била структура от тип асоциативен контейнер или някакъв тип дървовидна структура. И двата типа структури позволяват бърза актуализация, като за всяка нова дума се генерира автоматично шаблон, който се вмъква в контейнера или дървото и правила за образуване на словоформите, като шаблонът съхранява указател към тези правила.

Морфологичният анализатор разделя всяка въведена дума на морфеми, загатващи граматичните характеристики на думата и я сравнява с лексикалната база данни. По принцип всеки език има различно ниво на морфологична сложност и представя различни морфологични явления. Затова обикновено морфологичните анализатори за всеки език са специфични, използват новаторски методи и технологии, основани на характеристиките на конкретния език.

Има, разбира се, много учени-лингвисти, които се опитват да стандартизират нещата. Аврам Ноам Чомски създава йерархия от класове – формални граматики, образуващи формални езици. По-късно Мартин Кей и Роналд Каплан представят един модел, който теоретично може да се използва за морфологичното описание на който и да е език. Този модел е базиран на автомати с ограничено състояние (FSA = Finite State Automata) и преобразуватели с ограничено състояние (FST – Finite State Transducer). За най-голямо постижение в областта на компютърната морфология се счита моделът на Киммо Коскениеми, който реализира един морфологичен анализатор за финландския език по начин, изглеждащ възможен да се приложи за всеки друг език. Неговият метод подобрява модела на Кей и Каплан и е известен като морфология на две нива (two-level morphology). Тези модели са използвани и доразвити за новогръцкия език от Елени Галиоту и Анжела Рейли.

ЛИТЕРАТУРА

E. L. Antworth, PC-KIMMO: A Two-Level Processor for Morphological Analysis [Occasional Publications in Academic Computing 16], Summer Institute of Linguistics, Dallas TX,1990.

R. Kaplan and M. Kay. "Phonological rules and finite state transducers", in Proceedings of the ACL/LSA Conference, NewYork, 1981.

K. Koskenniemi, Two-level Morphology: A General computational Model for Word-Form Recognition and Production, University of Helsinki, 1983.

A. Ralli and E. Galiotou. "A prototype for a computational analysis of Modern Greek compounds", Asymmetry Conference, Université de Québec, à Montréal, May 2001.

Хр. Крушков. "Автоматизирано изграждане на машинни речници", XXV пролетна конференция на СМБ, 6–9 април 1996 г. Казанлък, с.199–204.