НОВИ ПОДХОДИ И ИНОВАЦИИ
NEW APPROACHES AND INNOVATIONS

АВТОМАТИЧЕН МОРФОЛОГИЧЕН АНАЛИЗ. ЕЛЕКТРОНЕН МОРФОЛОГИЧЕН РЕЧНИК

Елисавета Балабанова
Университет по библиотекознание и информационни технолигии

Резюме: *Накратко е представен въпросът за автоматичния морфологичен анализ. Разгледани са въпросите за това какво представлява морфологията и кои са основните операции в нея, както и какво представлява морфологичният речник в своя електронен вариант. Описано е как електронният морфологичен речник се използва в процеса на автоматичния морфологичен анализ. Описана е същността на автоматичния морфологичен анализ, както и най-често срещаните проблеми при неговото извършване. Накратко са представени процесите по лематизация и нормализация на текста.*

Ключови думи: *автоматичен морфологичен анализ; електронен морфологичен речник; лематизация и нормализация*

В тази статия е представен процесът по извършване на автоматичен морфологичен анализ. Автоматичният морфологичен анализ е една от първите стъпки в обработката на естествен език. Обработката на естествен език е част от научно-приложната област на изкуствения интелект, като специално в областта на книгоиздаването може да кажем, че технологиите на обработка на естествен език успешно се прилагат при редица практически операции, като програмите за правописна корекция, използването на всякакви видове електронни справочници и електронни речници, създаване на XML/HTML бази данни с продукцията на дадено издателство и др.

Морфология. Основни единици

Морфологията изучава правилата за образуване на думите и свързаните с различните форми на думата граматични значения.

Правилата за образуване на думите са езиково специфични. Построяването на нови думи в съответния език става с помощта на формообразуващите морфеми.

Морфемата е най-малката съставка на думата, която носи значение.

Пример: за - мин - ава - х - м - е
коренът **мин**- носи основното значение (срв. **мин**-а-ва-м, из-**мин**-ат)
представката *за*- модифицира значението (ср. *от*-минавам)
-х- променя глаголното време (срв. заминава*Х*ме и заминаваме).

Морфемите биват следните видове:

- *корен* – главната морфема; носи основното значение;
  Пример:
  корен *тек (кореновата гласна може да варира)
  ТЕК-а – глагол, сег. вр., несв. вид, 1 л. ед.ч.
  из - ТИЧ - ам – глагол, сег. вр., несв. вид, 1 л., ед.ч.
  раз - ТОЧ - в- ам – глагол, сег. вр., несв. вид, 1 л. ед.ч.
  ТОК – съществително, м.р., ед.ч.

- *представка* – морфема, която стои пред корена или пред друга представка;
  Пример:
  ПРИ-рода, ЗА-минавахме, РАЗ-ПРО-странявам

- *наставка* – морфема, която стои след корена или след други наставки;
  Пример:
  вод-ЕН

- *окончание* – морфема, която стои в края на думата и променя нейната форма.
  Примери:
  шофьор-И
  няколко стол-А

Морфологична категория система от противопоставени един на друг редове словоформи с еднородно съдържание.

Пример:
Морфологичната категория *род* в българския език се състои от три члена.

мъжки, женски, среден

хубав хубава хубаво

Морфологичната категория *число* в българския език се състои от два члена.

единствено, множествено

хубав, -а, -о; хубави

Отделните словоформи в даден лексикален ред на една морфологична категория се различават единствено по отношение на съответната си граматична характеристика.

Традиционно (тъй като езикознанието е доста стара наука) морфологията групира думите в няколко основни категории, наречени части на речта. В различните езици обемът и броят на морфологичните категории варират. В българския език частите на речта са десет: глагол, съществително име, прилагателно име, числително име, местоимение, наречие, междуметие, съюз, предлог, частица.

Основни операции в морфологията

В морфологията протичат две основни операции: словообразуване и формообразуване.

При *словообразуването* от една дума чрез добавяне на словообразуващи морфеми се образува/т нова дума/и. Изходната дума и новообразуваната се различават както по своето лексикално, така и по своето граматично значение.

Пример:
пиша -> писа-тел, раз-пис-ка, на-/пре-пиша

За разлика от словообразуването, при *формообразуването* от една дума чрез добавяне на формообразуващи морфеми се получава нова форма на същата дума. Двете думи имат едно и също лексикално значение, но тяхното граматично значение се различава.

Примери:
стол – столА
капя – капНА

За да изясним какви са процесите, извършвани при автоматичния морфологичен анализ, е необходимо да дефинираме какво представлява думата – както с оглед на традиционните дялове на езикознанието (лексикологията и лексикографията), така и с оглед на автоматичната обработка на езика.

Ортографската дума е последователност от букви между два интервала.

Лексемата е думата с всички свои граматични форми и значения.
Пример:

лексема: обичам
форми: обичам
обичаш
обичаме
ще обичам(е), щях(ме) да обичам(е), обичах, съм обичал и др.

В процеса на чънкиране при автоматичната обработка на езика се маркират първоначално границите на ортографските думи и на фразите. След това, в етапа на автоматичния морфологичен анализ, се пристъпва към определяне на границите на лексемите и свързването на дадена лексема със съответната нейна форма в даденото срещане в текста. За целта при последната операция (ако не се използват статистически методи) се използва електронен морфологичен речник.

Морфологичен речник

За да обясним същността на автоматичния морфологичен анализ (в чиято основа често е морфологичният речник), е необходимо да посочим какво представлява морфологичният речник.

Морфологичният речник в своя книжен вариант е представителна съвкупност от лексемите (основните форми на думите) на даден език, придружени с техните граматични характеристики. В морфологичния речник няма: тълкуване на думата, стилистични бележки, дефиниции, примери и друга лексикографска информация, която обикновено се дава в тълковните и в двуезичните речници.

В най-опростената си форма традиционният морфологичен речник е списък от думи с прикрепени към тях таблици, които оформят модела за формообразуването на всяка единица.

Компютърен морфологичен речник
Предназначенията на компютърния морфологичен речник са:

- за крайния потребител, т.е. за пряко ползване – дава бързо богата гама от справки върху морфологичните характеристики и формите на дадена лексема;
- като част от компютърна система, скрит за потребителя – тогава изпълнява същинската си роля – да установява връзките между формите на дадена дума, които се срещат в даден текст, и нейната основна форма, т.е. прави връзката *словоформа – лексема*.

Пример:
Лесно можем да изведем формата за 3 л. ед.ч. на глаголите в английския език от тяхната основна форма (*brings* < *bring*), но не така стои въпросът с формата за минало свършено време (т.нар. на английски past participle) – в този случай връзката не е така явна (*brought* < *bring* ?).

По същия начин стои и въпросът с извеждането на формата за множествено число на някои съществителни за мъжки род в българския език, където се наблюдава палатализация на съгласната в окончанието (*съпрузи* < *съпруг*), или пък на формата за миналото свършено деятелно причастие от глагол от свършен вид в минало свършено време (*отишъл* < *отидох*).

Двете роли на компютърния морфологичен речник определят и неговото приложение.

Приложения на компютърния морфологичен речник:

1) като езиковедски ресурс – в компютърния морфологичен речник се дава пълното словообразуване; дават се различни срезове на езика на нивото на неговата морфология;
2) като част от компютърна система.

Втората функция на речника е свързана с потребността в компютърната лингвистика да се обработват бързо големи масиви от текстове (напр. при всички информационни системи, в която и да е област – право, администрация и др.; при търсачките в интернет; в синтезиращия компонент на системите за машинен превод и т.н.).

Благодарение на компютърния морфологичен речник се извършва идентификация на буквените вериги – думите от текста – като членове на един и същи клас (т.е. като форми на една дума) и същевременно се прави минималната задължителна (граматична) обработка на езиковите единици и тяхната класификация.

Автоматичен морфологичен анализ

С помощта на компютърния морфологичен речник предварително токънизираният текст (за процеса на токънизация вж. например Simov et al. 2004 : 81–82) бива подложен на автоматичен морфологичен анализ. Възможностите за извършване на морфологичен анализ са както чрез статистически методи, така и с помощта на морфологичен речник. Тук разглеждаме само втората възможност.

При автоматичния морфологичен анализ се извършват две основни операции (Paskaleva 2007 : 2):

1) идентификация на езиковата единица;
2) класификацията ѝ.

Пример:
чел – минало свършено деятелно причастие от *чета*
1 -> *идентификация:* речникът открива мястото на тази форма сред 52-те форми на *чета* – мин. св. деят. прич., м.р., ед.ч.
2 -> *класификация:* речникът посочва, че тази форма унаследява характеристиките на глагола *несвършен вид* и *непреходност* и посочва свойствата на вида и родовата принадлежност.

При автоматичния морфологичен анализ:
- наборът от признаци, приписани на дадена езикова единица, се нарича граматичният маркер на думата, или *таг* (на англ. tag);
- тагът е резултат от пълния морфологичен анализ;
- процедурата на този анализ се нарича граматично маркиране или *тагиране* (tagging);
- минималната информация, кодирана в тага, е информацията за част на речта;
- таговете имат различна дължина и съдържание в зависимост от езика и от концепцията на съответната обработка на текст.

Всяка изследователска група, която обработва морфологично съответен езиков ресурс, решава какви да бъдат нивата на кодировка в таговете, които ще използва. Списъкът с тагове се оформя като списък от тагове (tag set) (вж. например решенията за Българския корпус със синтактични описания в: Simov, Osenova 2004).

Пример от Bultreebank – корпуса със синтактични описания на българския език:
Кабинетът (Ncmsf) ще (Vt---f-r3s) поиска (Vppt+f-r3s) от Европейския (Amsh) съюз (Ncmsi) да (C) излезе (Vppi+f-r3s) с декларация (Nsfsi) .
Ncmf – таг за съществително (noun) нарицателно (common), мъжки род (masculine), членувано (definite)

Проблеми при автоматичния морфологичен анализ

Някои елементи на текста не може да бъдат разпознати при автоматичния морфологичен анализ. Това са най-често (но не само): абревиатури (например СБА, УАСГ и т.н.); комбинации от числа и букви (1-вия, 23-то); комбинация от абревиатура и пълнозначна дума (ABS система); топоними (имена на държави, градове, реки и т.н.); лични имена и др.

Решението в тези случаи е да се съставят списъци с неразпознатите елементи. След това на всеки елемент от списъка лингвистите приписват ръчно адекватния според тях анализ и впоследствие изготвените списъци „се пускат“ върху вече морфологично анализиран текст. По този начин се покриват неразпознатите от първичния морфологичния анализ елементи.

Нормализация и лематизация на текста

Нормализация

След като текстът бъде морфологично анализиран, следва неговото нормализиране, или иначе казано:

1) Посочване, че две думи са една и съща;
   Пример:
   USA = U.S.A.

2) Премахване на интервалите на неправилни места вътре в думите/изразите;
   Пример:
   Те започнаха собсъждане на проектозакона. = Те започнаха с обсъждане на проектозакона.

3) Нормализиране на главните и малките букви.
   Пример:
   Докато той се прибираше, данчо му звънна. = Докато той се прибираше, Данчо му звънна.

Лематизация

След като бъде извършена нормализацията на текста, идва ред на следващата обработка, наречена лематизация. Лематизацията представлява свеждане на всички форми на дадена дума до основната ѝ форма. Този процес се извършва с помощта на морфологичния речник.

Примери:
съм, бях, ще бъда, бил съм → **съм** (това е лемата на всички форми отляво на стрелката)
дъжд, дъждове, дъжда, дъждът → **дъжд** (това е лемата на всички форми отляво на стрелката)

След извършването на лематизацията и при наличието на предварително подготвен списък с тагове се пристъпва към автоматичния морфологичен анализ, извършен с помощта на програма, наречена *тагер* (която може да бъде част от по-голяма компютърна система за обработка на естествен език и за създаване на електронни езикови ресурси).

Заключение

В настоящия текст беше разгледан накратко въпросът за автоматичния морфологичен анализ, който е в помощ на обработката на естествен език и създаването на електронни езикови ресурси, а в частност – и в помощ на книгоиздателския процес.

ЛИТЕРАТУРА
Паскалева, Е. (2007). *Компютърна морфология. Ресурси и инструменти*. София: ИПОИ БАН, ISBN 978-954-92148-1-9.
Simov, K., P. **Osenova** (2004). BTB- TR 04 BulTreeBank Morphosyntactic Annotation Bulgarian Texts. – In: *BulTreeBank* Technical Report BTB- TR 04. online at: http://bultreebank.org/bg/publications/
Simov, K. et al. (2004). Creation of a Tagged Corpus for Less-Processed Languages with CLaRK System**.** – In: *Proceedings of SALTMIL Workshop at LREC 2004: First Steps in Language Documentation for Minority Languages*, Lisbon, Portugal.

REFERENCES
Paskaleva, E. (2007). *Kompjutarna morfologia. Resursi i intstrumenti.* Sofia: IPOI BAS, ISBN 978-954-92148-1-9.

TAGGING, MORPHOLOGICAL DICTIONARY IN ELECTRONIC FORM, LEMMATIZATION AND NORMALIZATION OF TEXTS

Abstract: *The paper deals with the process of automatic morphological analysis. The issues of morphology and the operations in it are briefly discussed. The morphological dictionary in electronic form is discussed and also the process of automatic morphological analysis in NLP. The problems, encountered when doing automatic morphological analysis, are reviewed. The processes of automatic normalization and lemmatization are also discussed briefly.*

Keywords: *automatic morphological dictionary; automatic morphological analysis; normalization and lemmatization*

Assist. Prof. Elisaveta Balabanova, PhD
University of Library Studies and Information Technologies
119, Tsarigradsko shose Blvd.
1784 Sofia, Bulgaria
E-mail: e.balabanova @unibit.bg