КОМПЮТЪРНА ОБРАБОТКА НА ЕСТЕСТВЕНИ ЕЗИЦИ – ФОРМАЛИЗМИ ЗА АВТОМАТИЧЕН МОРФОЛОГИЧЕН АНАЛИЗ

Румен Рикевски

1. Съвременни тенденции в автоматичния морфологичен анализ

Едно от модерните направления на информатиката е компютърната лингвистика. Теоретичната компютърна лингвистика се занимава с формалните теории за лингвистичните знания, от които се нуждаят хората, за да генерират и използват езици. Днес тези достижения се базират изключително на компютрите.

Смисловата обработка на естествения език (която включва анализ и синтез) първоначално се свързва със синтаксиса и синтактичния анализ. Скоро обаче се установява, че оставането само на синтактично равнище не дава задоволителни резултати, защото понякога прилагането на едно или друго синтактично правило зависи от смисъла на самия текст. Появява се необходимостта от синтактично-смислов анализ (разбиране) на текста – извличане на знанията от текста и представянето им в някаква система. При решаването на този проблем се налагат редица ограничения (например на предметната област и на знанията за нея, на допустимите конструкции в езика, на обема на анализираната информация и т.н.).

При съвременната компютърна обработка на естествен език се изпълняват следните стъпки: морфологичен анализ, синтактичен анализ, семантичен анализ, интегриране в съответния контекст, прагматичен анализ.

При морфологичния анализ се анализират морфемите на думите, т.е. техните компоненти и признаци, а не участващите в думите признаци (например препинателните знаци) се отделят от думите.

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

Морфологичният анализатор представлява неразривна част от модерните системи за обработка на естествени езици и е непосредствено звено към речниковата база.

Съществуват два различни типа топологии на свързване на морфологичния анализатор към системите за обработка на естествени езици. При първия вариант системата търси в една речникова база, която съдържа пълните форми (full-form), за всяка една дума, която трябва да анализира. Ако думата не бъде намерена, тогава морфологичният анализатор поема функцията да направи предположение относно граматическите ѝ характеристики, базирано на нейната форма. При втората конфигурация морфологичният анализатор действа като посредник между системата и речниковата база. Всяка дума първо се разделя на съставните си морфеми и след това се търси в речниковата база, която е морфологична и съдържа морфеми – основи и окончания. Тази морфологична речниковата база е доста по-малка от речниковата база с пълните форми и затова тя е за предпочитане. Напр. за въвеждането на думата γιατρός в базата с пълните форми се изискват 8 записа γιατρός, γιατρού, γιατρό, γιατρέ, γιατροί, γιατρών, γιατρούς, γιατροί. Ако след това въведем думата αριθμός ще имаме още 8 записа. В морфологичната речникова база тези две думи ще се въведат по следния начин: първо въвеждаме 8-те окончания като морфеми (ος, ου, ο, ε, οι, ων, ους, οι); след това на тази група даваме един показател (таг) и го отбелязваме примерно с индекса A1; и накрая записваме думата γιατρός като (γιατρ + A1), а думата αριθμός като (αριθμ + A1), т.е. заявяваме, че двете думи се скланят по един и същи начин. По този начин речниковата база става много по-икономична, без да се губи никаква информация. При прилагателните имена печалбата е още по-голяма понеже напр. καλός -η -ο се скланя като διπλός -η -ο и в трите рода, а най-голяма е печалбата при глаголите. Използването на морфологичната речникова база е от решително значение при системи с голям речников потенциал, особено когато естественият език за обработка е с богата морфология както новогръцкия.

2. Модел Кей-Каплан

Когато става въпрос да се съединят две или повече морфеми при създаването на една словоформа, обикновено се проявяват някакви езикови феномени, които променят правописа на началните морфеми. Напр. когато трябва да се образува аорист на глагола δένω се събират основата δέν заедно с окончанието σα и аугмента ε, т.е ε-δέν-σα. За да се получи правилната словоформа έδεσα, ударението трябва да се премести една сричка вдясно и да се отнеме последното ν от основата. Този вид промени се описват от конкретни фонологични правила с помощта на които сме в състояние да извършим правилното преобразуване от една словоформа в друга. Представянето на една словоформа като поредица от морфеми, преди да се извършат необходимите промени (ε-δέν-σα) се нарича с термина лексикално представяне (lexical representation). Правилната словоформа на думата, която се образува след прилагането на фонологичните правила (έδεσα) се нарича с термина повърхностно представяне (surface representation).

Мартин Кей и Рон Каплан от компанията Xerox успели да представят всеки един фонологичен закон чрез т.нар. преобразувател на ограничено състояние (finite state transducer), който изразява последователното прилагане на правила и поетапното преобразуване на една дума от нейното лексикално представяне към повърхностното й представяне и обратно. Всеки един преобразувател действа между две състояния (представяния) на думата: формите преди и след прилагането на правилата. Последователните преходни състояния на думата (N-1 на брой, където N е броят на приложените правила) формират междинните състояния на словоформата. На N-тото ниво, след като сме приложили всички правила се образува повърхностното състояние на словоформата. Тази система за преобразуване на ограничено състояние прилага цялостната съвкупност от правила и може да действа в двете направления.

Моделът на Кей-Каплан изглежда способен да реши проблемите на морфологичния анализатор по един особено елегантен начин. На практика обаче, в езици с богата морфология, т.е. които трябва да приложат голям брой правила, мрежата която се получава е доста раздута и трудно се прилагат алгоритми за оптимизация. Въпреки това идеята за употреба на система, която извършва морфологични промени в двете направления става основа за създаването на преобразуватели (transducers) на състоянията на дадена дума.

3. Моделът на двете нива

Морфологичният модел на двете нива (two-level morphology model) е развит от Кимо Коскиениеми. Той представлява развитие на модела Кей-Каплан с преимуществото, че преобразувателите на ограничено състояние не са толкова големи. Коскиениеми използва своя модел във финландския език, който е с много богата морфология. Както показва и самото наименование, в модела на двете нива няма междинни състояния (представяния) на думата, както е в модела Кей-Каплан. Съществуват само лексикалното и повърхностното представяния и преобразуванията стават директно от едното състояние в другото и обратно. И тук всяко едно правило се изразява чрез преобразувател на ограничено състояние, но този път правилата не действат последователно, а паралелно.

При този модел правилата имат малко по-различна форма. Понеже се прилагат паралелно всяко правило трябва да се съобрази и с функционирането на съвкупността като цяло. Правилата са по-скоро забранителни отколкото заключителни както при модела Кей-Каплан. Всяко правило действа като филтър, забранявайки проявата на определени феномени. По този начин крайната форма е тази, която не е била възпрепятствана от нито едно от правилата.

Правилата на модела на двете нива разглежда всяка дума като съответствие на символи между лексикалното и повърхностното представяния. Напр. нека една дума има лексикално представяне mami, и повърхностно представяне mari. Всяка двойка от символи се нарича кореспондентна (m:m, a:a, m:r, i:i). Съответствията биват нормални, когато двата символа са еднакви (m:m) и специални, когато са различни (m:r). Всички възможни съответствия формират т.нар. разрешени двойки. В по-горния измислен пример, нека причината поради която второто m се превръща в r (m:r), а не в m, да бъде наличието на i което следва, т.е. нека в тази измислена граматика има правило, което казва „m се превръща в r когато следва i". Едно правило на модела на двете нива се състои от три части: кореспондентна двойка, оператор и обкръжение. Използват се 4 типа оператори: ограничение, наложено от контекста (context restriction), при който съответствието се проявява само в конкретно обкръжение; принуждение, породено от повърхностната форма (surface coercion), при който съответствието винаги се проявява в конкретното обкръжение; съставен оператор (composite), при който

съответствието се проявява винаги и само в конкретното обкръжение; изключение (exclusion), при който съответствието никога не се проявява в съответното обкръжение. Обкръжението определя фонологичната среда, в която се проявява съответствието.

Морфологичният модел на двете нива дава възможност като цяло да се развие една паралелна процедура за преобразуване с помощта на подходящия хардуер. Понеже преобразуването се базира на преобразуватели на ограничено състояние, то структурата може да поддържа както анализ така и синтез. Към крайния преобразувател, който е сравнително малък по-обем и сложност, може да се приложат актуалните правила на конкретната граматика.

ЛИТЕРАТУРА

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

2. **Allen, J. F.** Natural Language Understanding. The Benjamin/Cummings Publishing Company, 1987.

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

4. **Mackridge, P.** The Modern Greek Language. Oxford University Press, 1985.

5. **Τριανταφυλλίδ, Μ.** Νεοελληνιή Γραμματική, ΟΕΔΒ, 1986.

6. **Τριανταφυλλίδ, Μ.** Συντακτικό της Νέας Ελληνικής, ΟΕΔΒ, 1992.