Денисюк О. В.,
студентка ІІІ курсу,
Бондаренко О. О.,
студент І курсу
факультету української філології
Науковий керівник: Денисюк В. В. ,
Уманський державний педагогічний університет імені Павла Тичини

МОРФОЛОГІЧНИЙ АНАЛІЗ У СИСТЕМІ АВТОМАТИЧНОГО АНАЛІЗУ ТЕКСТУ

Автоматичний аналіз тексту включає ряд дуже складних операцій, які комп'ютер виконує з текстом на природній людській мові відповідно до алгоритму. При автоматичному аналізі текст послідовно перетвориться в його лексико-морфологічні, синтаксичні та семантичні уявлення, зрозумілі комп'ютеру. Зворотний процес перетворення лексико-морфологічних, синтаксичних і семантичних комп'ютерних уявлень у текст на природній мові називається автоматичним синтезом тексту.

Автоматичний аналіз і синтез тексту є важливими завданнями комп'ютерної лінгвістики як з точки зору розвитку теорії (розробки лінгвістичних основ створення штучного інтелекту), так і з точки зору реалізації практичних потреб людини, наприклад створення ефективних систем машинного перекладу.

Автоматичний аналіз тексту включає ряд етапів: а) *графемний аналіз*: виділення меж слів, речень, абзаців та інших елементів тексту (наприклад, *лід* «короткий виклад журналістського матеріалу, що розміщується після заголовку й перед основним текстом» у газетному тексті); б) *морфологічний аналіз*: визначення вихідної форми кожного використаного в тексті слова і набору морфологічних характеристик цього слова; в) *синтаксичний аналіз*: виявлення граматичної структури речень

тексту; г) *семантичний аналіз*: визначення змісту фраз. Ми торкнемося тільки особливостей морфологічного аналізу.

При морфологічному аналізі кожне використане в тексті слово зводиться до його вихідної форми, визначається набір морфологічних характеристик текстової форми слова: частина мови; рід, число і відмінок для іменників, число й особа для дієслів тощо. Кожне вжите в тексті слово називається словоформою (або слововживанням). Для забезпечення зв'язності тексту потрібний повтор тих самих слів, тому нерідко різні словоформи одного або декількох речень тексту зводяться до однієї і тієї ж вихідної формі, пор.:

Осінній день березами почавсь.

Різьбить печаль свої дереворити.

Я думаю про тебе весь мій час.

Але про це не треба говорити.

Ти прийдеш знов. Ми будемо на «Ви».

Чи ж неповторне можна повторити?

В моїх очах свій сум перепливи.

Але про це не треба говорити.

Хай буде так, як я собі велю.

Свій будень серця будемо творити.

Я Вас люблю. О як я Вас люблю!

Але про це не треба говорити (Л. Костенко).

Алфавітно-частотний словник словоформ цього вірша такий: осінній – 1, день – 1, березами – 1, почавсь – 1, різьбить – 1, печаль – 1, свої – 1, дереворити – 1, я – 4, думаю – 1, про – 4, тебе – 1, весь – 1, мій – 1, час – 1, але – 3, це – 3, не – 3, треба – 3, говорити – 3, ти – 1, прийдеш – 1, знов – 1, ми – 1, будемо – 2, на – 1, ви – 1, чи – 1, ж – 1, неповторне – 1, можна – 1, повторити – 1, в – 1, моїх – 1, очах – 1, свій – 2, сум – 1,

перепливи – 1, хай – 1, буде – 1, так – 1, як – 2, собі – 1, велю – 1, будень – 1, серця – 1, творити – 1, вас – 2, люблю – 2, о – 1. Крім незмінних службових (*але*, *про*, *не*, *як*), повнозначних частин мови (*я*, *це*, *треба*, *говорити*, *люблю*), ужитих 2–4 рази, укажемо на присвійний займенник *свій*, ужитий у формах називного відмінка однини *свій* і знахідного відмінка множини *свої*; присвійний займенник *мій*, ужитий у формах знахідного відмінка однини *мій* та місцевому відмінку множини *моїх*; особовий займенник 2 особи множини *ви*, ужитий у формах називного відмінка *ви* і знахідного відмінка *вас*; дієслово *бути*, ужите у формі 1 особи множини простого майбутнього часу *будемо* та 3 особи однини наказового способу *хай буде*.

У словниках зазвичай перераховано не словоформи, а слова, зведені до певної вихідної форми. Вихідною формою використаних у тексті словоформ залежно від типу мови може слугувати *лема* (словникова форма лексеми) або *основа* (ядерна частина слова без словозмінних морфем). Наприклад, українські словоформи *робити*, *робитиму*, *роблю*, *роблячи*, *роблено*, *робивши* мають одну лему *роб*.

Флективним і аглютинативним мовам з багатою словозміною для збереження всіх можливих словоформ потрібні значні ресурси пам’яті. Наприклад, український іменник, що змінюється за числами (2 числа) і відмінками (7 відмінків), має 14 словоформ. Українське дієслово характеризується складнішим набором граматичних характеристик і, відповідно, має більшу кількість словоформ. У цьому випадку за вихідну форму, до якої зводиться слово, доцільніше брати його основу.

Однак у морфологічному аналізі термін «основа» не завжди має зміст, укладений у нього в канонічній граматиці. Наприклад, якщо у слові засвідчено чергування (*сидіти – сиджу, друг – друзі* та ін.), то основою (точніше, квазіосновою, або машинною основою) в цих випадках виступає

частина слова не тільки без словозмінних морфем, а й без букв на позначення звуків, що чергуються, тобто *си#* і *дру#*, відповідно. Такий тип виокремлення основ отримав назву стемінг, тобто зведення різних словоформ до однієї квазіоснови. Стемінг повністю задовольняє вирішення деяких автоматичних завдань, наприклад пошук в Інтернеті. Так, запиту користувача *фотографи* повній або неповній квазіоснові відповідають іменник *фотографія* і прикметник *фотографічний*. У результаті пошуку користувач отримає список документів зі словосполученнями *фотографічний портрет*, *портретна фотографія*.

Для морфологічного аналізу важливе не тільки поняття машинної основи, яке трактують як послідовність літер від початку словоформи, спільну для всіх словоформ, що входять у формоутворювальну парадигму слова. Наступний крок – це визначення частиномовної належності слова (частиномовний тегінг) і його морфологічних характеристик, що найчастіше відбувається з опорою на словозмінні елементи слова (машинні закінчення).

Машинні закінчення – елементи, що описують формозміну конкретної лексеми і подані у вигляді парадигм. Усі можливі набори машинних закінчень зафіксовані в типовій парадигмі лексеми. При цьому, з одного боку, можна спостерігати збіги типових парадигм (і, відповідно, машинних закінчень) різних лексем, наприклад *ручка* і *онучка*, а з іншого, збіги машинних основ лексем, що мають різні типові парадигми, пор. типові парадигми машинної основи *мат#*, що належать іменнику *мати* і дієслову *мати*.

За машинними закінченнями, що входять у певні типові парадигми, здійснюється повна морфологічна характеристика кожної словоформи, напр.:

Я {я = Р, ос. = Н. в., одн.}

купив {купити = V, док. = мин., одн., дійсн., чол., перех.}

мила {мило = S, серед., неіст. = Р. в., одн. | = S, серед., неіст. = Н. в., мн. | = S, серед., неіст. = Зн. в., мн. | мити = V, недок. = мин., одн., дійсн., жін., перех. | мила = А, жін., Н. в., одн. | = S, іст., жін., Н. в., одн.}

щоб {щоб = С}

моя {моя = Р, присв. = Н. в., одн.}

мила {мила = S, іст., жін., Н. в., одн. | = А, жін., Н. в., одн. | мило = S, серед., неіст. = Р. в., одн. | = S, серед., неіст. = Н. в., мн. | = S, серед., неіст. = Зн. в., мн. | мити = V, недок. = мин., одн., дійсн., жін., перех.}

мене {я = Р, ос. = З. в., одн.}

мила {мити = V, недок. = мин., одн., дійсн., жін., перех. | мила = А, жін., Н. в., одн. | = S, іст., жін., Н. в., одн. | мило = S, серед., неіст. = Р. в., одн. | = S, серед., неіст. = Н. в., мн. | = S, серед., неіст. = Зн. в., мн.}

У наведеному аналізі можна побачити лексико-морфологічну багатозначність третього, шостого і восьмого слова. Вибір правильної форми здійснюється людиною з урахуванням синтаксичної ролі слова в реченні та його значення. Автоматичний дозвіл багатозначності або зняття омонімії – вибір правильної інтерпретації словоформи, що допускає кілька варіантів тлумачень, – відбувається шляхом ручної розмітки або автоматично, на основі ймовірнісних моделей (наприклад, в англійській мові найбільш імовірне поєднання невизначеного артикля з іменником) або на основі правил, створених автоматично чи людиною.

Для автоматичного морфологічного аналізу використовують парсери – спеціальні комп’ютерні програми для автоматичного аналізу слів.

Отже, для досягнення максимально повного ефекту морфологічний аналіз має складатися з таких етапів:

1) нормалізація словоформ, що має вигляд лематизації, тобто зведення

різних словоформ до єдиного представлення – до вихідної форми (леми) або стемінгу, тобто зведення різних словоформ до однієї квазіоснови;

2) частиномовний тегінг, тобто вказівка частини мови для кожної словоформи в тексті;
3) повний морфологічний аналіз – приписування граматичних характеристик словоформі.