УДК 004.82 (045)

СТРУКТУРА ЛІНГВІСТИЧНОГО ПРОЦЕСОРУ СИСТЕМИ ПОРІВНЯЛЬНОГО АНАЛІЗУ ТЕКСТІВ ЗА ЗМІСТОМ

А.І. Вавіленкова
Національний авіаційний університет
e-mail: a_vavilenkova@mail.ru

Труднощі автоматичної формалізації текстових документів полягають головним чином у різноманітті форм природної мови, за допомогою яких може виражатися одна й та сама інформація, наприклад, абревіатури, синоніми, омоніми, інверсний порядок слів у реченні, не однозначний запис числівників, вставні слова, вигуки, авторські знаки та ін.

Формалізація текстової інформації передбачає побудову функціональної моделі природної мови, якою може виступати логіко-лінгвістична модель речень, як транслятор, що встановлює відповідність між текстом та його змістом. Це означає посилення ролі формальної логіки, оскільки у випадку формальних теорій вже не можна задовольнятися інтуїтивними твердженнями, так як та чи інша аргументація узгоджена з логічними правилами, набутими завдяки здатності до правильного мислення.

Обробка знань в комп'ютері представляє собою обробку їх змісту правилами перетворення тих форм, якими описуються знання в машині. Тому при обробці знань фундаментальною і важливою проблемою є опис змістового вмісту проблем широкого діапазону, а також наявність опису такої форми знань, яка гарантуватиме, що їх обробка формальними правилами перетворення буде здійснюватися правильно.

Однорідне представлення приводить до спрощення механізму управління логічним виводом та спрощення управління знаннями. Показником інтелектуальності системи, з точки зору представлення знань, вважається здатність системи використовувати в потрібний момент необхідні (релевантні) знання. В проблемі доступу до знань можна виділити три аспекти:

1) зв'язність (агрегація) знань є основним способом, що забезпечує прискорення пошуку релевантних знань; знання організовують навколо найбільш важливих об'єктів (сутностей) предметної області;

2) механізм доступу до знань – деякий опис сутності, призначений для відшукання в базі знань об'єктів, що відповідають цьому опису; очевидно, що впорядкування та структурування знань можуть значно пришвидшити процес пошуку;

3) способи співставлення: синтаксичні, параметричні, семантичні та примусові.

У випадку синтаксичного співставлення співвідносяться форми (зразки), а не зміст об'єктів. Успішним вважається співставлення, в результаті якого зразки ідентичні. Результат синтаксичного співставлення є бінарним. У параметричному співставленні вводиться параметр, що визначає ступінь співставлення. У випадку семантичного співставлення співвідносяться не зразки об'єктів, а їх функції.

Компонента системи, що реалізує створення формальної лінгвістичної моделі та здатна працювати з природною мовою у всьому її об'ємі, є *лінгвістичним процесором*. Дві основні функції лінгвістичного процесору (ЛП) полягають у вилученні змісту із заданої текстової інформації та вираженні отриманого змісту мовою логіки предикатів для можливості подальшого порівняння сформованих моделей [1]. Зміст висловлювання – це вся семантико-прагматична інформація, яку користувач передає на вхід системи. Внутрішнє представлення змісту містить сутності проблемної області (слова), що потрапляють до системи з даним висловлюванням, властивості та відношення, що відповідають цим сутностям.

Лінгвістичний процесор представляє собою багаторівневий перетворювач, що містить три рівні представлення тексту: морфологічний, синтаксичний та семантичний (рис. 1).

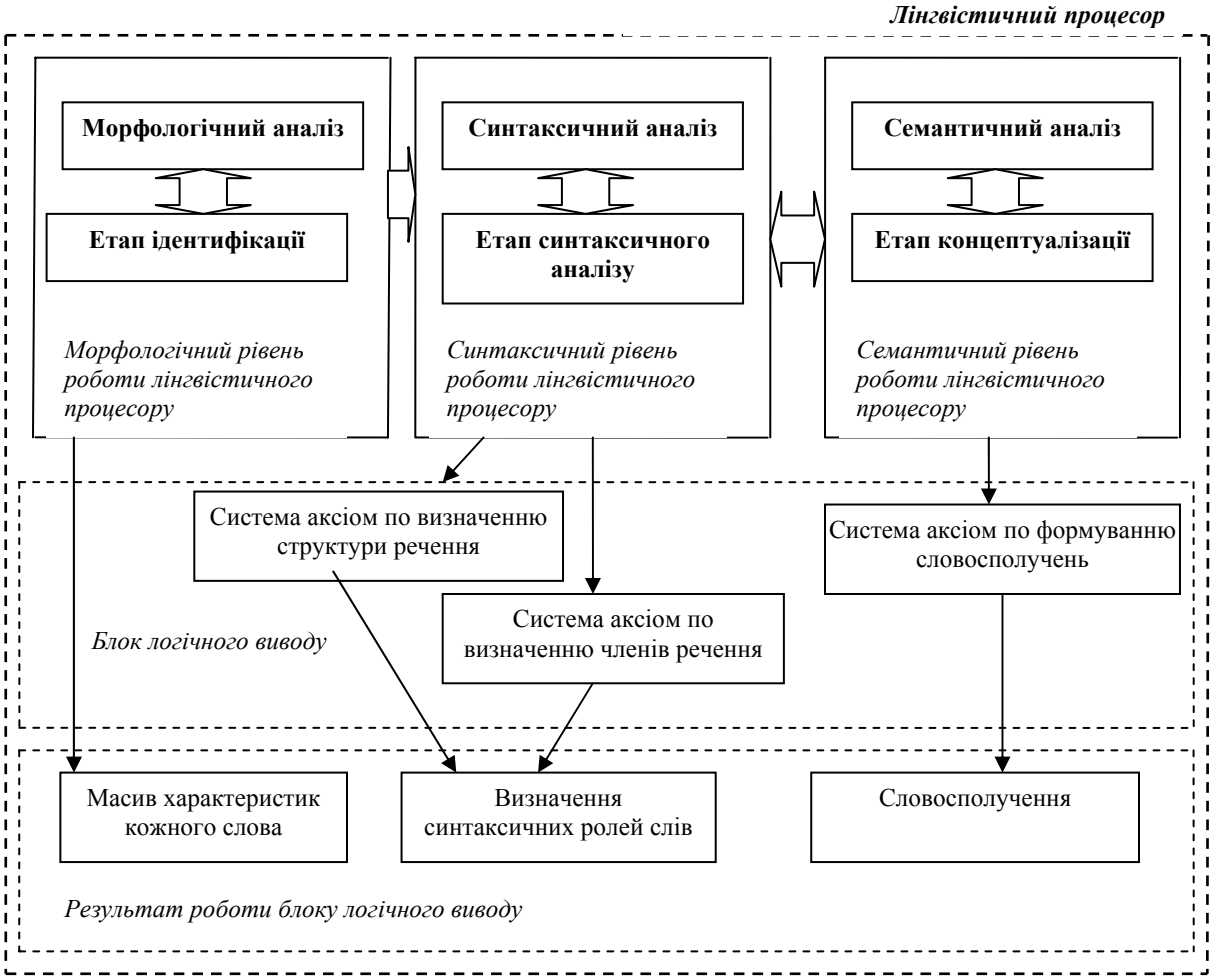

Рис. 1. Лінгвістичний процесор (рівні та структура)

Існує твердження, що повністю можуть бути формалізовані лише елементарні теорії з простою логічною структурою і невеликим запасом понять. Розробка лінгвістичного процесору за схемою рис. 1 спростує це твердження і доведе, що формалізувати можна довільну текстову інформацію, представлену у вигляді різних типів речень. Детальніший розгляд кожного рівня лінгвістичного процесору допоможе це зрозуміти.

Метою **морфологічного аналізу** є визначення граматичних характеристик словоформ для використання їх на наступних етапах обробки тексту. Існує три основних методи реалізації морфологічного аналізу [2].

1) *Декларативний* – в словнику зберігаються всі можливі словоформи кожного слова з прописаною для них морфологічною інформацією, у цьому випадку задача морфологічного аналізу полягає у пошуку словоформи в словнику та прописування їй відповідних характеристик, тому в цьому методі відсутній сам морфологічний аналіз, а зберігається тільки його результат. Перевагою цього методу є висока швидкість аналізу, а також універсальність по відношенню до словоформ, проте він не вирішує омонімії флективної мови. Декларативний метод доцільно застосовувати при частому зверненні до лінгвістичного процесора.

2) *Процедурний* – виділяє в кожній конкретній словоформі основу, ідентифікує її та приписує даній основу відповідний для неї комплекс морфологічної інформації. Метод

передбачає попередню систематизацію морфологічних знань про флективну мову та розробку алгоритмів отриманої інформації конкретній словоформі. Недоліком такого підходу є висока трудомісткість складання словників сумісності, а наявність в природній мові великої кількості слів-винятків не дозволяє повністю автоматизувати цей процес. Реалізація процедурного методу займає невеликий об'єм пам'яті, проте збільшує час пошуку. Ще одним недоліком є відсутність універсальності, так як існує велика кількість слів, які не можливо представити у вигляді суми незмінної основи та афіксів.

3) *Комбінований* – використовує як словник словоформ, так і словник основ. На першому етапі відбувається пошук по словнику словоформ, як при декларативному методі, і у випадку успішного пошуку аналіз на цьому завершується. В протилежному випадку використовується словник основ і процедурний метод аналізу. За рахунок використання готових словоформ забезпечується досить висока швидкість визначення граматичних форм, а за рахунок словника основ та афіксів – якість морфологічного аналізу.

Згідно з денотативною концепцією розуміння значення речення, зміст текстової інформації полягає у визначенні відношень між висловлюваннями та екстралінгвістичною ситуацією (подією), що описується цим висловлюванням. Тобто аналізується співвідношення синтаксичних функцій членів речення і тих ролей, які вони виконують. Комбінування слів в реченні залежить від їх ролі, а також від характеру того синтаксичного зв'язку, який їх з'єднує.

Основною задачею семантичних пошуків в області синтаксису речення вважається з'ясування значення формально-синтаксичних моделей. Згідно цієї точки зору граматична організація речення вже сама по собі являється фактором, вагомим для семантичної структури побудованого по цій схемі речення. Значення компонентів цієї схеми слугують основою семантичної структури речення, представляючи її в максимально загальному вигляді. У відповідності з цим під семантичною структурою речення розуміється його інформативний зміст, представлений в абстрагованому вигляді, закріплений у мовній системі відношень, типізованих елементів змісту. Отже, основою роботи систем обробки текстової інформації являється синтаксичний та семантичний аналізатори [3].

Синтаксичний аналіз – це процес, який визначає, чи належить деяка послідовність лексем мові, що породжується граматикою. По будь-якій граматиці можна побудувати синтаксичний аналізатор, але граматики, які використовуються на практиці, мають спеціальну форму.

Синтаксичний розбір речення відбувається шляхом набору послідовних перетворень:

- пошук граматичних ідіом;
- лексико-граматичний аналіз речення з усуненням неоднозначності у визначенні частин мови;
- знаходження іменної групи об'єкта і суб'єкта;
- знаходження дієслівної групи;
- виділення головних та залежних речень.

Саме на цьому етапі, коли визначені синтаксичні ролі структурних елементів речень, що розглядаються, можна здійснити їх порівняльний аналіз, розглянувши побудовані логіко-лінгвістичні моделі. Нехай логіко-лінгвістичні моделі двох розглянутих речень мають вигляд:

$$P_1 = (x_1, x_2[x_3\{c_{31}\}[x_4[x_5\{c_{51}\}]]], x_6[x_7][x_8[x_9[x_{10}], x_{11}[x_{12}[x_{13}]]]]),$$
$$P_2 = (, x_2, [x_3\{c_{31}\}[x_4[x_5\{c_{51}\}]]], x_6[x_7][x_8[x_9[x_{10}], x_{11}[x_{12}[x_{13}]]]]),$$

де P – предикат, що відображає зміст речення; x_1 – предикатна змінна (суб'єкт), знаходиться у предикативному відношенні з P; c_{1d_1} – предикатна константа, що вказує на

ознаку суб'єкта; d_1 - номер предикатної константи, що вказує на ознаку суб'єкта; x_q - предикатна змінна (аргумент); q - номер предикатної змінної (аргументу), початкове значення якого $q = 2$; c_{qd_2} - предикатна константа, що вказує на ознаку q-тої предикатної змінної (аргументу або об'єкта); d_2 - номер предикатної константи, що вказує на ознаку предикатної змінної (аргументу);

Порівняння логіко-лінгвістичних моделей заданих речень відбувається за такими принципами:

1. порівнюються предикати речень P_1 та P_2: якщо P_1 та P_2 – дієслова, їх час співпадає, P_2 – дієслово в пасивній формі, а P_1 – дієслово в активній формі, а також співпадають корені та суфікси предикатів, або P_1 та P_2 є синонімами, то предикати P_1 і P_2 можна вважати тотожними за змістом;

2. якщо в одному із речень, що порівнюються, відсутній суб'єкт, а x_1 та x_2 іменники у називному та знахідному відмінку відповідно, а кількість предикатних змінних у одному з речень на одну більше, ніж у іншому, x_1 та x_2 – спільнокореневі слова або є синонімами, то суб'єкти речень (віртуальний у випадку пасивного стану) можна вважати тотожними за змістом;

3. якщо справджуються умови 1 та 2, то речення можна вважати однаковими за змістом.

Семантичний аналіз тексту базується на результатах синтаксичного аналізу, отримуючи на вході уже не набір слів, розбитих на речення, а набір дерев, що відображають синтаксичну структуру кожного речення. Так як методи синтаксичного аналізу поки що не дають бажаних результатів, рішення цілого ряду задач семантичної обробки тексту базується на результатах аналізу окремих слів, і замість синтаксичної структури речення аналізуються набори слів, що стоять поряд.

Загальною базою методів семантичного аналізу, яка дозволяє виявити семантичні відношення між словами, є тезаурус мови. На математичному рівні він представляє собою орієнтований граф, вузлами якого є слова в їх основній словоформі, а дуги задають відношення між словами і можуть відображати ряд ознак. Таким чином тезаурус задає набір бінарних відношень на множині слів природної мови. Такі тезауруси для російської та англійської мови знаходяться на стадії розробки та являються закритими ресурсами. Семантичний аналіз тексту включає в себе ряд практичних задач. Однією з них є вилучення інформації з текстів та представлення її у вигляді формальної системи знань. У цьому напрямку виконано ряд експериментальних розробок, проте готових програмних продуктів немає. Подібні системи існують і призначені для інтеграції та очистки даних, що знаходяться в сховищах, але вони не надають ніяких засобів вводу даних.

Складність реалізації високоточного аналізатора пов'язана з тісним зв'язком між синтаксисом та семантикою, наявністю в текстах великої кількості синтаксичних омонімічних конструкцій, які не допускають однозначної інтерпретації без залучення знань про семантичну сумісність слів. Формалізована модель мови призначена об'єднати синтаксичну та семантичну складову. Не дивлячись на обмежену точність синтаксичних аналізаторів, їх використання здатне помітно підвищити якість таких систем у випадку комбінування з статистичними методами.

Література

1. *Осуга С.* Обработка знаний/ Пер. с япон. – М.: Мир, 1989. – 293 с.

2. *Хант Э.* Искусственный интеллект / Хант Э. – М.: Мир, 1978. – 560 с.

3. *Поспелов Д. А.* Логико-лингвистические модели в системах управления / Поспелов Д.А. – М.: Энергоиздат., 1981. – 232 с.