Автоматизація процесу вилучення знань з електронних документів

К.т.н., доцент ВАВІЛЕНКОВА АНАСТАСІЯ ІГОРІВНА
Національний авіаційний університет, м. Київ, Україна

Існуючі сьогодні засоби аналітичної обробки електронних документів не здатні здійснювати глибокий лінгвістичний аналіз текстової інформації. Для цього необхідні інструменти, що функціонуватимуть на основі принципів розуміння природної мови. Особливо актуальною ця проблема є у сфері освіти, науки, законотворчості, патентування, інноваційній діяльності, у площині професійних інтересів різних юридичних та інформаційно-аналітичних організацій і підрозділів, пов'язаних із захистом інформації, де з'являється потреба у визначенні збігів або виявленні логічних протиріч у текстових документах. А отже, існує необхідність створення сучасних засобів автоматичного вилучення знань з електронних документів на основі побудови формальних моделей та алгоритмічної бази формалізованого опису структур природної мови.

Для того, щоб простежити зв'язки між окремими об'єктами в реченні природної мови, суб'єктами та їх діями, пропонується використати схему автоматизованого формування логіко-лінгвістичних моделей текстової інформації, що базується на синтаксичному парсингу.

Синтаксичний парсинг – це процес співставлення лінійної послідовності лексем мови з його формальною граматикою. Фактична реалізація схеми автоматизованого формування логіко-лінгвістичних моделей текстової інформації розуміє під собою сортування масиву слів конкретного речення у відповідності з чітко сформованими правилами. Основною ідеєю цього перетворення є визначення відношень між підметом (суб'єктом) та присудком (предикатом). Вважається, що синтаксис формалізовано, якщо існують правила, які виражають граматичне відношення між членами речення та дають можливість встановити зміст окремих структурних одиниць тексту. Це можливо шляхом використання системи продукційних правил формування словосполучень природної мови.

Схема автоматизованого вилучення знань з речення природної мови включає в себе декілька етапів.

1. Ідентифікація характеристик кожного слова природно-мовного речення. На вхід системи подається речення природної мови. Програма розбиває його на окремі слова. Для кожного слова, що потрапляє на вхід системи, утворюється окремий клас:

public class Words implements IDeclarations {//клас, що відповідає простому елементу системи

Words leftNeighbour = null;// елемент системи, що знаходиться зліва

Words rightNeighbour = null; // елемент, що знаходиться справа
private Znak zn = null; // змінна, що відповідає за знак пунктуації
private int column; // номер стовпчика в таблиці бази даних
*/** Властивості простого елемента */*
private StringBuffer name = new StringBuffer(); // Слово
private int cm = 0; // частина мови
private int g = 0; // відмінок
private int n = 1; // число
private int k2 = 0; // рід
private int t = 0; // час
private int h = 0; // спосіб
private int l = 0; // особа
private int ch = 0; // член речення}

Тобто для подальшого розбору під кожне слово виділяється вісім характеристик:

$$Z_i = \{cm_{i1}, g_{i2}, n_{i3}, k2_{i4}, t_{i5}, h_{i6}, l_{i7}, ch_{i8}\},$$

де $cm_{i1}, i1 = \overline{1,11}$ – граматична характеристика, що позначає частину мови, може приймати одне із значень: cm_1 – іменник; cm_2 – прикметник; cm_3 – числівник; cm_4 – займенник; cm_5 – дієслово; cm_6 – дієприкметник; cm_7 – дієприслівник; cm_8 – прислівник; cm_9 – прийменник; cm_{10} – сполучник; cm_{11} – частка; $g_{i2}, i2 = \overline{1,7}$ - морфологічна ознака, що відповідає за відмінок; $n_{i3}, i3 = \overline{1,2}$ - граматичний параметр, що означає число; $k2_{i4}, i4 = \overline{1,4}$ - граматичний параметр, що означає рід; $t_{i5}, i5 = \overline{1,3}$ - граматичний параметр, що означає час; $h_{i6}, i6 = \overline{1,3}$ - граматичний параметр, що означає спосіб; $l_{i7}, i7 = \overline{1,3}$ - граматичний параметр, що означає особу. $ch_{18}, i8 = \overline{1,5}$ - означає член речення, яким виступає слово.

Таким чином, кожне слово речення, що подається на вхід системи характеризується набором цифр типу: *{1,1,1,0,0,3,1,0}, {2,1,1,0,0,3,3,1}, {3,1,1,0,0,3,3,1}, {4,1,1,0,0,3,3,1}, {5,1,1,0,0,3,3,1}, {6,1,1,0,0,3,3,1}, {1,2,1,0,0,3,1,0},{2,2,1,0,0,3,3,1},{3,2,1,0,0,3,3,1}, {4,2,1,0,0,3,3,1}, {5,2,1,0,0,3,3,1}, {6,2,1,0,0,3,3,1}.* Наприклад, характеристики Z_i = 1,1,1,0,0,3,1,0 означають, що слово S_i – іменник у називному відмінку, чоловічого роду однини, третьої особи, підмет.

2. Виявлення зв'язків між словами речення природної мови

Для всіх флективних форм характерно, що наступний за словом рівень синтаксичних конструкцій – це словосполучення. У кожній природній мові слова у словосполученнях пов'язані за певними законами, наприклад, в українській мові – відношення прилягання, узгодження, керування. Аналізуючи ці відношення, можна сформулювати ряд правил, за якими неявно формуються словосполучення, та формалізувати їх, базуючись на тому, що слова пов'язані за рахунок знаходження та

узгодження певних граматичних ознак. Всі сформульовані правила об'єднаємо у систему продукцій.

Нехай є словосполучення $S_j =$ «технічний словник», де «технічний» – це слово S_i, а «словник» – слово S_{i+1}. Це словосполучення можливо сформувати через відповідність граматичних форм кожного із слів:

$$if\ (cm_i == 2)\ \&\&(cm_{i+1} == 1)\ \&\&(g_i == g_{i+1})\ \&\&(n_i == n_{i+1})\ \&\&(k_i == k_{i+1})\ \&\&(l_i == l_{i+1})$$
$$then\ (S_j = S_i \cup S_{i+1})$$

Правило трактується наступним чином: якщо частина мови для першого слова прикметник, а для другого – іменник, відмінок, число, рід та особа обох слів співпадає, то слова утворюють словосполучення.

За аналогічною схемою складаються правила формування всіх словосполучень природної мови. Здійснюється аналіз всіх можливих взаємозв'язків між різними частинами мови, а не між конкретними словами. Для української мови сформована система продукцій, що містить тридцять одне правило формування словосполучень, тридцять два правила визначення синтаксичних ролей та тридцять правил визначення типів речень.

3. Формування логіко-лінгвістичної моделі речення природної мови. Керуючись виявленими зв'язками та отриманими характеристиками кожного зі слів, будується логіко-лінгвістична модель, яка представляє собою одновимірний масив, значеннями елементів якого є слова речення, упорядковані згідно з формулою:

$$P(x_1\{\underset{d_1 \in C_1(x_1)}{\wedge} c_{1d_1}\}, \underset{q1 \in J(S)}{\wedge}[\underset{q \in X(S)}{\wedge}[x_q\{\underset{d_2 \in C_2(x_q)}{\wedge} c_{qd_2}\}]]),$$

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; x_q – предикатна змінна (аргумент); d_2 – номер предикатної константи, що вказує на ознаку предикатної змінної (аргументу).

Наприклад, для речення *«Предметом інформатики являється досить швидка обробка інформації за відомими законами»* логіко-лінгвістична модель буде мати такий вигляд:

$$P(x_1\{c_1\_c_{11}\}[x_4[x_5\{c_{51}\}]], x_2[x_3]),$$
$$являється(обробка\{швидка\_досить\}[інформації[законами\{відомими\}]],$$
$$предметом[інформатики]).$$

Задача формалізації текстової інформації вимагає одночасного об'єднання зусиль спеціалістів в області інформаційних технологій та вчених у сфері лінгвістики. З точки зору прагматики можна говорити про те, що предметом синтаксису та семантики є граматична структура речення, що і намагався довести автор за допомогою досліджень, описаних у даному матеріалі.