

ОСНОВНІ БЛОКИ КОМП'ЮТЕРНОЇ СИСТЕМИ ЛІНГВІСТИЧНОГО АНАЛІЗУ ЕЛЕКТРОННИХ ДОКУМЕНТІВ

*канд. техн. наук, доц. А.І. Вавіленкова, Національний авіаційний
університет, м. Київ*

Комп'ютерна система лінгвістичного аналізу призначена для автоматизованої змістовної обробки електронних документів. В основі її функціонування лежить формальний апарат побудови логіко-лінгвістичних моделей текстової інформації. До складу системи входять такі блоки.

Блок введення текстової інформації передбачає використання інтерфейсу користувача для введення файлу у форматі *.txt*.

Блок членування тексту здійснює розбиття всього електронного документу, що потрапляє на вхід системи, на декілька рівнів. Внутрішнє представлення змісту повинно містити концепти предметної області, властивості та відношення, що приписуються цим концептам.

Лінгвістичний процесор системи здійснює перетворення природно мовного тексту в змістовну модель, що складається з множини логіко-лінгвістичних моделей речень.

Аналізатор логіко-лінгвістичних моделей речень природної мови використовує правила виявлення синонімічних конструкцій, а також правила застосування засобів когезії, таких як семантичне та дейктичне повторення, анафоричні зв'язки, дискурсивні слова, синтаксичний паралелізм та ін.

Компонента побудови змістовної моделі електронного документу. Внаслідок застосування правил у попередньому блоці системи змістовна модель зазнає змін, в результаті чого на виході системи отримуємо набір логіко-лінгвістичних моделей речень, що в сукупності дають змістовну модель електронного документу.

Автоматизований лінгвістичний аналіз текстів можна здійснити лише шляхом застосування формалізованих правил виявлення між фразових зв'язків. Таким чином, особливістю створеної системи є база знань, що представляє собою множину продукцій, кожна з яких описує правило вживання одного із засобів когезії. Саме формалізація таких правил дала змогу враховувати лексико-синтаксичні перетворення в електронних документах. Результати роботи системи лінгвістичного аналізу електронних документів можна екстраполювати на тексти різних стилів, внаслідок чого за уривками текстів складати уявлення про засоби зв'язку у всьому тексті.