

Автоматичний синтаксичний аналіз

Робота присвячена методам оцінки альтернативних моделей автоматичного аналізу тексту. Синтаксичний аналіз – це процес зіставлення лінійної послідовності словоформ мови з її формальною граматикою. Метою синтаксичного аналізу є розпізнання синтаксичної структури речення або окремих його частин. У процесі синтаксичного аналізу встановлюється, чи відповідає даний ланцюг словоформ правилам граматики мови. Автоматичний синтаксичний аналіз має на меті за допомогою алгоритмів отримати в явному вигляді синтаксичну структуру речення. При цьому під «реченням» розуміють частіше за все просте речення, тобто таке, в котрому як складників немає інших простих речень, а складне речення називають зазвичай «фразою» (Ю. Н. Марчук).

Основними проблемами моделювання синтаксису в прикладній лінгвістиці є подання синтаксичних структур природної мови й розробка оптимальних методів (алгоритмів) синтаксичного аналізу, орієнтованих на те чи те подання синтаксичної структури. Основні дії, які виконує синтаксичний компонент, – це аналіз, трансфер та синтез. Синтаксичний аналіз має на вході морфологічне представлення кожного слова, фрази, а на виході – синтаксичне представлення перекладеної вхідної фрази. Існують різні способи зображення синтаксичної структури речення, серед яких можна виокремити 3 основні та найбільш поширені: дужковий запис, зображення залежностей у вигляді стрілочок, спрямованих від головного слова до залежного, зображення синтаксичної структури у вигляді дерева (В. В. Іванов).

Зараз широко використовуються три основних способи подання синтаксичної структури речення: структури складових (дерево безпосередньо складових), структури залежностей (дерево залежностей) та структури, які поєднують властивості структур залежностей та складових (орієнтовані структури складових) (Л. В. Бондаренко). У моделі складових речення розглядається як лінійно упорядкований ланцюжок словоформ речення. На лінійно упорядкованому ланцюжку словоформ речення задається бінарне відношення залежності; так відображається змістовно той факт, що дві словоформи у реченні, які утворюють словосполучення, зазвичай «нерівноправні» як у граматичному, так і у змістовому плані. Це відношення залежності можна розглядати як об'єднання традиційно виокремлених у реченні відношень керування, узгодження, керування та прилягання.

Запропонований у роботі метод оцінки лінгвістичних моделей дозволяє оцінити якість алгоритмів синтаксичного аналізу на попередніх етапах їх розробки і не вимагає трудомісткого етапу програмування алгоритму. Також тут використовуються алгоритмічні, статистичні методи та метод моделювання. Ми проілюстрували принципи автоматичного синтаксичного аналізу тексту на прикладі алгоритму, який був розроблений Г. П. Александровим, Є. А. Загіком, А. К. Родіоновим та Є. І. Стоговим (Г. Г. Белоногов). Цей алгоритм виявляє поверхневу синтаксичну структуру речення, і при цьому на нього не накладається ніяких обмежень з точки зору словникового складу.

Аналіз тексту в розглядуваному алгоритмі проводиться за реченнями. Для кожного речення будується його дерево залежностей, в якому відображаються літерні коди слів, зв'язки між словами та граматична інформація до слів. У процесі аналізу фіксується лише факт існування синтаксичного зв'язку між словами та напрям цього зв'язку. Більш детальна диференціація зв'язків не здійснюється. Сурядний зв'язок розглядається як указівка на відсутність безпосереднього зв'язку між словами та словосполученнями та їх належність до одного й того ж елементарного тексту.

Робота складається з вступу, трьох розділів (розділ 1 – Поняття синтаксичного аналізу, розділ 2 – Спосіб подання синтаксичної структури, розділ 3 – Алгоритм синтаксичного аналізу за частинами мови) та висновків.