ГурєєваЛ.В.

Національний технічний університет України «КПІ», кафедра АМТС№1

ІНСТРУМЕНТАЛЬНІ ЗАСОБИ ЛІНГВІСТИЧНИХ ДОСЛІДЖЕНЬ

Однією з найважливіших задач штучного інтелекту (ШІ) є спрощення взаємодії людини з комп'ютером за рахунок створення ефективних способів спілкування на природній мові. Для реалізації даного завдання необхідно розробити програмний комплекс, що дозволяє розпізнавати будь-яке текстове повідомлення природної мови (аналіз) і/або кодувати його в певну формальну мову (переклад), або виконувати якусь дію згідно результату аналізу прийнятого повідомлення (діалог). Подібний програмний комплекс буде складатись з досить складних уніфікованих компонент, які призначені виконувати поетапне розпізнавання та обробку тексту за допомогою лінгвістичних методів, що використовують одну або декілька баз даних і словників. Найважливішими компонентами цього комплексу є продукційна компонента, необхідна для обробки продукційних правил, що визначаються лінгвістичними правилами, а так само словникова компонента, яка реалізує роботу з базами даних і словниками.

Багато прикладних лінгвістів в даний час [1,2] працюють над створенням систем побудови різних природно-мовних інтерфейсів, які орієнтовані на класифікацію текстів в структурованій певним чином предметній області, і на підтримку запитів до баз даних. Подібні системи використовують семантичний аналіз природної мови, де розуміння засноване на змістовної інформації, яка характеризує дану предметну область. Для вирішення цієї задачі необхідно розробити і реалізувати алгоритми морфологічного та синтаксичного компонентів лінгвістичного процесора природної мови (ПМ). Лінгвістичний процесор (ЛП) являє собою комплекс програм, що забезпечують аналіз і синтез тексту на природній мові. Завданням такого процесора є розбір і обробкаінформації, що надходить у ЛП у вигляді окремих фраз на природній мові (при аналізі) або побудова фрази ПМ, яка відповідає формальному опису її сенсу (при синтезі).

Вкрай важливим завданням ШІ є побудова семантичної мережі. Найчастіше для її опису використовують концептуальні графи Дж. Соува і блокові структури Г. Хендрікса. Блокові структури докладно описані в [3].

Перші спроби інтелектуальної обробки текстів на природній мові були зроблені в 60х- 70х роках. З цією метою було створено багато експериментальних програм, які здатні "спілкуватися" з користувачем на природній мові. Для цього використовувалися концепції грамотного і структурного програмування. Грамотне програмування (ГП, Literate Programming), яке іноді помилково називають літературним програмуванням, являє собою методологію програмування і документування. Це словосполучення іноді помилково перекладають як «літературне програмування». Цей термін і концепцію в 1981 році висунув Дональд Кнут. Грубо кажучи, ГП - це мова програмування, написане на «псевдокоді» - «людській мові». Текст програми на такій мові зрозумілий і легко сприймається програмістами, в той же час сам код важкий, а під однією фразою - «оператором» ховається безліч інших вкладених абстракцій або програмний код безпосередньо машинною мовою. Система грамотного програмування, яку Кнут запропонував як альтернативу «структурному програмуванню» в 1970 -х роках, незважаючи на доведену ефективність, мало поширена сьогодні через нерозуміння: багато хто думає , що ГП - це всього лише система документування або форматування стандартних коментарів [4].

В цілому, програми, що використовують лінгвістичний процесор, поки не отрималиширокого розповсюдження. Найчастіше причиною було невисока якість розпізнавання фраз, жорсткі вимоги до синтаксису «природної мови», великі часові та ресурсні витрати, необхідні для якісного функціонування. Майже у всіх системах комп'ютерної обробки тексту використовується обмежена природна мова, оскільки поки не створено повної формальної моделі для жодної природної мови.

Видається за доцільне формувати моделі великих фрагментів ПМ (статті, книги тощо) у вигляді двох зв'язаних таблиць реляційної бази даних швидкохідних СУБД типу ORACLE. Рядки першої таблиці містять повний опис

вершин (вузлів) семантичної мережі у вигляді об'єкта і його характеристик. Наприклад, в якості об'єкта можуть виступати слова або пропозиції, а характеристиками є їх різні ознаки і додаткова інформація, яка використовується різними програмами для семантичного аналізу. В якості опорних об'єктів (вузлів), виступають слова, які використовуютьсяу заголовку або анотації конкретного фрагмента ПМ. Ця таблиця за своєю структурою подібна словнику, має гнучку можливість довільної деталізації збережених об'єктів.

Друга таблиця описує зв'язки вузлів першої таблиці у вигляді базових смислових фрагментів, лінгвістичних правил і призначена для вирішення завдань аналізу, перекладу чи діалогу за допомогою навченої бази даних.

На закінчення необхідно відзначити, що природні мовні засоби спілкування комп'ютера і людини продовжують розвиватися і це є одним з найбільш перспективних способів, який дозволить створити складні інформаційні системи для спілкування людини і машини на природній мові.

Література:

1. Волкова И.А. «Лингвистический процессор естественного языка.Морфологический и синтаксический компоненты. Задание практикума для студентов 3-го курса ЧФ МГУ (Методическое пособие)».Издательский отдел факультета вычислительной математики и кибернетики МГУ им. М.В.Ломоносова, 2002. -39 с.
2. Волкова И.А., Руденко Т.В. Формальные грамматики и языки. Элементы теории трансляции. — М., Изд-во МГУ, 1999.
3. Hendrix G.G. Encoding Knowledge in Partitioned Networks // Associative Networks: Representations and Use of Knowledge by Computers. – New York: Academic Press, 1979.

Інтернет-ресурс:
http://ru.wikipedia.org/wiki/%D0%93%D1%80%D0%B0%D0%BC%D0%BE%D1%82%D0%BD%D0%BE%D0%B5_%D0%BF%D1%80%D0%BE%D0%B3%D1%80%D0%B0%D0%BC%D0%BC%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5