ЩОДО ПРОГРАМИ СИНТАКСИЧНОГО КОРЕКТОРА

В.Ю. Шелепов, А.Н. Лимарь, А.В. Шевченко, Г.В. Саввіна,
Э.Н. Селищев, В.А. Грабовая
Інститут проблем штучного інтелекту
340048, Донецьк, Артема, 118-б,
Тел.: (0622) 92-6082, Факс: (622) 92-6082,
E-Mail:shel@iai.donetsk.ua

Computer russian speech recognition is complicated by great number of wordforms. It is appropriate the computer should identify close wordforms and restore right syntax using special program, which is not connected with recognition.The such kind program is the subject of the report. It is based on wordform construction libreries for more then 100 thousands words in initial forms, which were written by authors.

ВСТУП

Проблему комп'ютерного розпізнавання усної мови на тій чи іншій природній мові, напевно, необхідно роздивлятися як найважливійшу частину більш загальної задачі навчання комп'ютера відповідній мові. При цьому ситуація до деякої міри нагадує навчання людини іноземній мові. Але для викладача очевидно, що знання мови - це не тільки знання або розуміння слів, але й володіння мовними структурами. Серед останніх одне з найважливіших місць займають синтаксичні структури.

Російська мова, якою ми займаємося, відноситься до флективних мов й відрізняється великою кількістю форм слів. Тому для неї проблеми синтаксису пов'язані з задачею розпізнавання мови. Справа в тому, що для комп'ютера дуже важко відрізняти близькі за звучанням форми слів, наприклад, «садит - садят» або «экспонента - экспоненты». Тому бажано використовувати програми, які дозволяли б комп'ютеру ототожнити такі словоформи одного й того ж слова. Тоді усна мова буде розпізнаватися у вигляді ланцюжка слів у найбільш зручних для розпізнавання формах. Але при цьому виникає задача відновлення у введенному тексті правильного синтаксису, яка повинна вирішуватися окремою комп'ютерною програмою, не пов'язанною з розпізнаванням - синтаксичним коректором. Такий коректор буде принципово видрізнятися від програм, які використовуються в автоматичних перекладачах, тому що тут відсутній первісний синтаксично правильний текст, й відновлення синтаксису необхідно здійснювати практично з нуля. Очевидно, що рішення цієї задачі у повній мірі неможливо без враховування семантики і, більш того, без участі людини. Ця участь в ряді моментів повинна відбуватися у режимі діалогу з комп'ютером. Авторами цього повідомлення розроблена програма, яка відновлює правильний синтаксис достатньо простого речення російської мови в ланцюжці слів, які знаходяться в початкових формах. Це ті форми слів, які знаходяться в словнику.

1. ПРО УТВОРЕННЯ СЛОВОФОРМ

В основі синтаксичного коректора, який описується, лежать написані авторами програми утворення словоформ для словника, який містить більш ніж 100 000 слів. Як джерело лексичної та синтаксичної інформації був використанний відомий «Граматичний словник російської мови» А.А.Зализняка [1] та його електронна форма. Словник містить розмітку слів та вказівки, як, орієнтуючись на неї, будувати всі необхідні словоформи. Відповідні алгоритми зводяться в основному до відкидання закінчення, перетворення основи при наявності біглої голосної й додавання нового закінчення. Іноді основа повинна повністю замінюватися іншою (наприклад, «идти» - «шел»).

Трудність полягає у великій кількості різновидів конкретних реалізацій цих алгоритмів. Так, для відмінювання іменників використовуються близько двохсот різновидів алгоритму утворення п'яти непрямих відмінків однини й шести відмінків множини. Розмітка значної частини словника була нами змінена з причин, про які сказано нижче.

Метод, який застосовується при утворюванні словоформ, розглянемо на прикладі іменників. Іменники російської мови, що містяться в граматичному словникові А.А.Зализняка, були розподілені на наступні групи:

- Група А. Іменники, при утворенні форм яких їх основа не змінюється, а змінюється лише закінчення в кінці слова;
- Група В. Іменники, при утворенні форм яких основа змінюється;
- Група С. Іменники цієї групи позначають або людей за національною, географічною та соціальною належністю (наприклад, «южанин») або малят (наприклад, «бельчонок»)

• Група D. Іменники цієї групи мають нестандартне закінчення «-а» в називному відмінку множини (наприклад, «рукава»), або нестандартне «нульове» закінчення в родовому відмінку множини (наприклад, «грузин»).

Для кожної з вищеозначених груп була складена структура, яка містить всі можливі закінчення, які виникають при утворенні відмінкових форм в однині та множині і написана функція підстановки цих закінчень.

Всі іменники зі складу словника А.А. Зализняка були перерозмічені, тому що розмітка автора словника не підходила для роботи з ряду причин.

Приклад 1. (приклад нової розмітки)

абажур mAa0_0

З цієї розмітки бачимо, що:

• слово «Абажур» має чоловічий рід. На це вказує літера 'm';

• слово «Абажур» - неістота. Це видно з того, що літера 'm' маленька. Якби слово вказувало на істоту, розмітка починалася б з великої літери;

• Закінчення додається в кінець слова без знищення останньої літери слова. На це вказує наявність в розмітці літери 'a'. В противному випадку в розмітці замість литери 'a' стояла б літера 'b'. (Наприклад, слово «портфель mAb2_1»);

• При підстановці закінчень однини використовується структура за номером '0', при підстановці закінчень множини - також структура за номером '0';

Як можна бачити з описання нової розмітки, на неї покладене практично все інформаційне навантаження про утворення форм слів, які відрізняються від початкових, і, отже, нова розмітка більш громіздка порівняно з розміткою А.А.Зализняка. І хоча, казалося б, словник, який містить слова з новою розміткою, повинен займати більше місця на диску, ніж вихідний, його розмір приблизно дорівнює розміру вихідного словника, тому що з останнього було викинуто безліч службової інформації, на яку в багатьох випадках спирається використання словника А.А.Зализняка.

Крім того, робота з новою розміткою дозволяє отримати значний виграш у часі на стадії утворення словоформ. Це можна продемонструвати на наступному прикладі.

Приклад 2. Розглянемо наступні два слова: слово «Площадь» и слово «Мышь». В словнику А.А.Зализняка вони мають однакову розмітку - «8е». Але як наслідок того, що основа другого слова закінчується на шиплячий, а основа першого - ні, в утворенні форм «площадям - мышам», «площадями - мышами» є різниця. Таким чином, утворюючи словоформи за розміткою А.А.Зализняка й зустрівши одне з вищезгаданих (або аналогічних) слів з розміткою «8е», ми змушені робити перевірку на наявність шиплячої в основі слова й в залежності від результату перевірки підставляти необхідне закінчення. Ці дії неможливо реалізувати одним або декількома операторами алгоритмічного язика. Крім того, оператори перевірки й порівняння, особливо строкових констант, займають одне з перших місць по захвату процесорного часу.

Ми розглянули випадок, коли тільки дві дещо відмінні групи слів віднесені до одного класу. В вихідному словнику зустрічаються випадки, коли до одного класу віднесені 4-5 груп іменників, які різняться між собою (див. [1], с. 48). В зв'язку з цим кількість перевірок на стадії утворення словоформ збільшується. Нова розмітка, як нам здається, вільна від подібних дефектів.

Утворення словоформ для інших самостійних частин мови здійснюється за аналогічними алгоритмами. Програми утворення словоформ оформлені у вигляді динамічних бібліотек (DLL), які можна підключати в готовому вигляді до проектуємих граматичних програм і які представляють, таким чином, самостійний програмний продукт.

2. ПРО ПРОГРАМУ КОРЕКТОРА

У нас вони використовуються в програмі «Синтаксичний коректор», робота якої базується на використанні апарату дерев синтаксичного підпорядкування (див. [2]). Після того, як ланцюжок початкових форм введено у вікно редактора, комп'ютер через діалогове вікно питає про спосіб і час, в які необхідно поставити дієслово, яке виражає присудок. Те ж саме він робить відносно числа, в яке необхідно поставити підмет (відповіді - натискування відповідних кнопок за допомогою миші). Після цього комп'ютер автоматично узгоджує підмет і присудок. У випадку наявності додатку він ставить іменник у відмінку, який визначається прийменником. Нарешті, він узгоджує наявні прикметники й числівники з тими іменниками, до яких вони відносяться, й ставить у потрібній формі займенники.

ВИСНОВКИ

В результаті ми отримуємо можливість обробити ланцюжок слів типу: "Усталый конь медленно пробирается сквозь высокий трава", перетворивши її в речення: «Усталые кони медленно пробирались сквозь высокую траву».

ЛІТЕРАТУРА

1. Зализняк А.А. Грамматический словарь русского языка. Москва: "Русский язык", 1977, 880с.

2. Гладкий А.В. Синтаксические структуры естественного языка в автоматизированных системах общения. Москва: Наука, 1985, 144с.