УДК 004.421

К.т.н. Заболотня Т. М., бакалавр Глушаускайте І.В.

Національний технічний університет України
«Київський політехнічний інститут»

ПОБУДОВА СЛОВНИКА ТА ТЕКСТОВОГО КОРПУСУ ДЛЯ ПРОГРАМНОГО ГЕНЕРАТОРА РИМОВАНОГО ТЕКСТУ

Abstract

Tetiana M. Zabolotnia, snr.teacher, PhD; Irina Glushauskaite, student
Creating a dictionary and a text corpora for a rhymed text generator

This article concerns creating a dictionary and a corpus of poetic texts for the rhymed text generator. The properties of the dictionary and the corpus to create are described. The most significant text corpora and some dictionaries with inter-word semantic relations are also described. At the end, some conclusions about the data to use and the way to present it are made.

Вступ

Автоматизована генерація римованого тексту на сьогоднішній день є однією з цікавих задач у галузі комп'ютерної обробки природномовних текстових даних. Розробка ефективного програмного забезпечення синтезу чи аналізу тексту завжди базувалась на використанні сучасних досягнень філологів та програмістів щодо створення комп'ютерних лінгвістичних ресурсів. Але побудова віршованого тексту через специфічність об'єкту обробки є оригінальною задачею, вирішення якої потребує використання спеціалізованих словників та баз даних. Таким чином, дослідження та розробка ресурсів для підтримки роботи програмного генератора римованих текстів є актуальною задачею.

Постановка задачі

Для коректної роботи генератор віршів слід забезпечити відповідними вхідними даними, такими як словник заданої мови, що може бути підключений до програми та містить необхідні для генерації тексту дані, та корпус семантично коректних римованих текстів, складених заданою мовою.

Звідси мета даної роботи полягає у зборі відомостей про існуючі словники, лінгвістичні бази даних та корпуси англомовних текстів (особливо віршованих текстів), аналізі можливостей їх використання та виборі з них найбільш придатних для розв'язання задачі генерації римованого тексту. Також необхідним є визначення структури програмного словника та способу його наповнення даними, потрібними для роботи програми-генератора.

Термінологія

Корпус або *корпус текстів* – великий структурований набір текстів [1].

Семантична мережа – інформаційна модель предметної галузі, що має вигляд орієнтованого графа, вершини якого відповідають об'єктам предметної галузі, а дуги (ребра) задають відношення між ними [1].

Тезаурус (у лінгвістиці) – особливий різновид словників загальної або спеціальної лексики, в яких вказані семантичні відношення (синоніми, антоніми, пароніми, гіпоніми, гіпероніми тощо) між лексичними одиницями [1].

Створення корпусу римованих текстів

Вимоги до корпусу. З текстів корпусу програма має отримувати статистичні дані щодо порядку слів у реченнях, сумісного вживання слів тощо. Оскільки програма генерації римованих текстів формує вірші, подібні до віршів використовуваного нею корпусу, останній повинен містити римовані тексти, складені сучасною мовою. Разом з тим, бажано щоб він містив якомога менше зразків футуристичної, модерністської поезії тощо через нестандартні слововживання та порядок слів у такій поезії.

Огляд існуючих корпусів. Коротко розглянемо найбільш широко використовувані корпуси англійських текстів.

- British National Corpus –корпус сучасної англійської мови, за взірцем якого створювалось багато сучасних корпусів різноманітних мов. Корпус містить більш 100 млн. слововживань, а також підкорпус усної мови на 10 млн. слововживань. Характеризується використанням повних текстів різного стилю та напрямку [2];
- Cambridge International Corpus– створювався перш за все як база для розробки навчальних матеріалів та словників англійської мови. Містить велику колекцію текстів. Недоліком даного корпусу є те, що на даний момент до нього мають доступ лише автори, що працюють над книгами для видавництва Cambridge University Press [3];

• The Bank of English – корпус англійської мови, що постійно поповнюється; містить 524 млн. слововживань. Надає можливість вибору підкорпусу: британські книги, журнали тощо; американські книги, радіопередачі тощо. На жаль, підкорпусу поезії тут немає. Крім того, доступ до повної версії корпусу є платним [3].

Слід зазначити, що існують також такі корпуси віршованих текстів як корпус англосаксонської поезії [4, 5], корпус російської поезії у англійському перекладі, корпус американської поезії 18-19 сторіч [6] тощо. Але, на жаль, за такими вузькоспеціалізованими текстами неможливо навчити програму сучасній поетичній мові.

Як бачимо, існуючі корпуси текстів, складених англійською мовою, здебільшого не містять віршованих творів, а деякі корпуси навіть не є вільно доступними. Тому у даній роботі пропонується скомпонувати поетичні тексти самостійно. Звідси постає нова задача – збір великої кількості римованих текстів, групування їх за певними ознаками (наприклад «адаптовані до сучасної мови вірші Шекспіра», «англійська поезія 19 сторіччя» тощо). У разі успішного вирішення даної задачі програма-генератор під час формування римованого тексту підтримуватиме додаткову опцію включення або виключення окремих баз текстів з процесу генерації тексту. Також за умови наявності декількох різнопланових корпусів текстів можливим буде створення стилізованих віршів, наприклад «генерація привітання у стилі Роберта Бернса».

Створення словника

Вимоги до словника. Перш, ніж аналізувати наявні словники, визначимо, яким вимогам повинен відповідати результуючий словник:

- словник повинен містити відомості про частину мови для кожного слова;
- бажано, щоб у словнику була подана інформація про наголос у слові;
- словник повинен містити інформацію про семантичні зв’язки між словами, причому як загальновживані (синоніми, антоніми, пароніми тощо), так і введені нами. Для забезпечення цієї властивості потрібно проаналізувати зібраний раніше корпус текстів. Для подання цих зв’язків зручним буде зберігання словника у вигляді семантичної мережі.
- словник повинен мати зручний інтерфейс для підключення його до створюваної програми. Такою формою може бути сукупність серіалізованих об’єктів класу «слово», до кожного з яких задані посилання на слова, пов’язані з цим словом.

Аналіз наявних ресурсів. Проаналізуємо існуючі словники з точки зору їх відповідності зазначеним вище вимогам.

- WordNet – семантична мережа слів англійської мови, розроблена у Прінстонському університеті [7] . Вузлами мережи є «синсети» - множини синонімів, що мають спільний сенс, та список коротких загальних визначень для цих слів. Між синсетами існують зв'язки – «синоніми», «гіпоніми-гіпероніми» тощо.

- Вікісловник – багатофункціональний багатомовний словник і тезаурус, один з проектів фонду «Вікімедіа». Це спроба поєднати граматичний, етимологічний, тлумачний словники, а також тезаурус. Містить інформацію про семантичні зв'язки слів.

Ці словники є зручними за структурою, але не мають механізмів інтеграції до створюваної програми. Проаналізувавши структуру цих словників, було обрано деякі їх риси, що стануть в нагоді для майбутньої реалізації словника.

Організація словника генератора римованого тексту. За структуру словника обрано семантичну мережу, яку пропонується реалізувати засобами ООП. Вузлами мережі є об'єкти класу «слово», ребрами – різноманітні зв'язки між словами. Існують різні типи зв'язків, наприклад «слово, якому безпосередньо передує дане слово», «слово, за яким безпосередньо слідує дане слово», «слово, що міститься з даним словом у одному реченні». Для подання цих зв'язків у кожному об'єкті класу «слово» необхідно задати масив посилань на слова, пов'язані з цим словом. Для кожного посилання також існує тип. У зв'язків повинна бути вага, вона відповідає кількості випадків або імовірності того, що ці слова зустрічаються у заданому порядку. Також вводяться зв'язки типу «слово-синонім», «слово-антонім» тощо. Словник у такому випадку є сукупністю серіалізованих об'єктів класу «слово». Для побудови мережі пропонується використати два словники - звичайний словник, що містить перелік слів із зазначенням частин мови, та словник синонімів, антонімів тощо.

Створений таким чином словник, що має структуру у вигляді семантичної мережі, є зручним для використання об'єктноорієнтованою програмою-генератором римованого тексту.

Висновки

Проведений аналіз існуючих мовних інформаційних ресурсів, які можуть бути використані у процесі генерації римованого тексту, таких як словники, лінгвістичні корпуси, семантичні мережі тощо. В процесі аналізу виявлені властивості, яким повинні відповідати вхідні дані для генератора римованого тексту. Словника і корпусу текстів, що відповідають цим властивостям, зараз не існує, тому було визначено доцільну структуру подання цих даних і спосіб їх накопичення.

Література

1. Вікіпедія [Електронний ресурс http://ru.wikipedia.org/], дата візиту 06.03.2010
2. British National Corpus [Електронний ресурс http://www.natcorp.ox.ac.uk/], дата візиту 07.03.2010
3. Сайт Російська Лінгвістика [Електронний ресурс http://rusling.narod.ru/qqq_corp_nonslav_engl.htm], дата візиту 07.03.2010
4. The Online Corpus of Old English Poetry (OCOEP) [Електронний ресурс http://www.oepoetry.ca/], дата візиту 07.03.10;
5. The Complete Corpus of Anglo-Saxon Poetry [Електронний ресурс http://www.std.com/obi/Anglo-Saxon/aspr/contents.html], дата візиту 07.03.10
6. American Verse Project [Електронний ресурс http://quod.lib.umich.edu/a/amverse/], дата візиту 07.03.2010
7. Офіційний сайт мережі WordNet [Електронний ресурс http://wordnet.princeton.edu/], дата візиту 07.03.2010
8. Розенталь Д. Э. и др. Словарь лингвистических терминов [Електронний ресурс http://www.gumer.info/bibliotek_Buks/Linguist/DicTermin/index.php], дата візиту 06.03.2010