УДК 004.421

К.т.н. Заболотня Т.М., бакалавр Ларіонов М.О.

Національний технічний університет України
«Київський політехнічний інститут»

ІТЕРАТИВНИЙ АЛГОРИТМ КВАЗІСЕМАНТИЧНОЇ ГЕНЕРАЦІЇ РИМОВАНОГО ТЕКСТУ

Abstract

Tetiana M. Zabolotnia, snr.teacher, PhD; Mykyta Larionov, student
An iterative algorithm of quasisemantic rhymed text generation

This article deals with the problem of generating quasisemantic rhymed text. A generation algorithm is based on iterative selection of new lines using a linguistic database in the form of a semantic network. This algorithm helps to solve the main problems in rhymed text generation.

Вступ

У наш час у зв'язку з виникненням великої кількості задач, пов'язаних з автоматичною обробкою і синтезом текстових даних, відчутного значення набуває проблема генерації тексту, близького до природної мови. Прикладом таких програм можуть бути людино-машинні інтерфейси і генератори спеціалізованих текстів, які стали важливою частиною ІТ-простору.

На даний момент не існує ефективних рішень задачі програмної генерації семантично коректного римованого тексту. Існуючі засоби представлені у вигляді генераторів тексту з деякої заданої множини готових рядків або генераторами наступного рядка по кільком заданим рядкам і ключовим словам. Це зумовлює актуальність розробки нового алгоритму автоматизованої генерації римованого тексту, що не має вищезазначених недоліків.

Постановка задачі

Задача полягає в створенні нового алгоритму генерації семантично коректного римованого тексту, який буде враховувати основні недоліки існуючих рішень, такі як низький рівень «змістовності», невелику кількість можливих варіантів згенерованих текстів та відсутність гнучкого механізму щодо задання вхідних даних.

Термінологія

Рима – звуковий збіг слів або їх частин в кінці ритмічної одиниці при відсутності змістового збігу [1].

Віршований розмір — правило чергування ненаголошених і наголошених складів у вірші [1].

Склад – голосний або інший довгий дзвінкий звук разом з одним і більше приголосними, передуючим йому або наступним за ним, які вимовляються як єдине ціле [1].

Наголос - виділення одного зі складів у складі слова різними фонетичними засобами [1].

Семантична мережа — інформаційна модель предметної галузі, що має вигляд орієнтованого графа, вершини якого відповідають об'єктам предметної галузі, а дуги задають зв’язки між ними [2].

Опис алгоритму

Алгоритм генерації квазісемантичного римованого тексту базується на роботі зі спеціалізованим словником, який містить такі лінгвістичні відомості про слова як:

1) слово;
2) дані про слово:
    а) частина мови;
    б) кількість складів;
    в) наголошений склад;
3) посилання на слова, які можуть стояти перед даним з вказаними вагами зв’язків, відповідно до ймовірності зустріти слово в тексті перед даним;
4) посилання на слова, які можуть стояти після даного з вказаними вагами зв’язків, відповідно до ймовірності зустріти слово в тексті після даного;
5) посилання на синоніми;
6) посилання на слова, які можуть бути в одному реченні з даним.

Таким чином, словник утворює собою лінгвістичну мережу, в якій всі слова пов’язані семантичними зв’язками, за рахунок використання яких підвищується імовірність формування змістовних словосполучень або фраз під час генерації римованого тексту.

Виконання алгоритму передбачає задання в якості вхідних даних певного шаблону тексту, який включає в себе такі елементи:

- кількість і вид строф (наприклад, дистих, терцет, катрен, квінтет, секстина, септіма, тощо);
- ритм і віршовий розмір;
- ключові слова з зазначенням їх позицій;
- словник (набір слів, які використовуються для генерації);
- правила римування рядків (вказуються номера рядків у строфі, які повинні римуватися або відсутність римування);

Існує можливість автоматичної генерації шаблону з готового вірша, аналізуючи його по кількості і виду строф, ритму і віршовому розміру, типу римування.

Римований текст складається зі строф (куплетів), які, в свою чергу, складаються з рядків. Розглянемо створення кожного елемента.

Кожен рядок генерується за таким принципом:

1. Якщо в рядку немає ключових слів, то ми вибираємо довільне слово зі словника і вважаємо його ключовим
2. Використовуючи семантичну мережу, рекурсивно знаходимо слова, які логічно пов'язані з ключовим, поки не отримаємо рядок потрібної довжини, або не буде перевищено деякий радіус пошуку по семантичній мережі. Оптимальна відстань між ключовим і кінцевим словом в семантичній мережі становить 2-3 слова.

Створення кожної строфи представляє собою послідовність таких кроків:

1. Створення першого рядка строфи відповідно до розміру.
2. Створення запиту для формування наступного рядка.
3. Створення масиву випадкових рядків відповідно до віршового розміру з урахуванням ключових слів.
4. Визначення релевантності рядків відносно запиту.
   а) визначаємо значення релевантності для рядка за такими критеріями:
      - відповідність до ритму;
      - відповідність до правил римування;
      - точність фонетичного збігу рими;
      - максимальна відстань між словами в семантичній сітці;
   б) якщо релевантність вище визначеного порогу – додаємо рядок до набору можливих варіантів;
   в) якщо масив випадкових рядків не вичерпано – переходимо виконуємо пп.а, б для наступного рядка;
   г) якщо набір можливих варіантів пустий, переходимо до кроку 3;
5. З набору можливих варіантів вибираємо найбільш релевантний і додаємо його до строфи

6. Якщо розміру строфи досягнуто – алгоритм закінчується, інакше переходимо до пункту 2.

Алгоритм генерації строфи повторюється до досягнення потрібної кількості строф відповідно до початкового шаблону тексту.

Висновки

Запропонований алгоритм генерації квазісемантичного римованого тексту дозволяє отримувати результат, який може з високою ймовірністю відповідати вимогам користувача щодо рівня «змістовності» вірша. Алгоритмом передбачено використання лінгвістичної бази даних у вигляді семантичної мережі та розширення набору вихідних даних шаблоном тексту.

Алгоритм здатний вирішити виділені проблеми таким чином:

- проблему низької семантичної зв'язності слів – за рахунок використання семантичної сітки під час синтезу фраз і словосполучень;
- проблему низької кількості варіантів вихідних текстів – за рахунок підключення і вибору потрібних словників;
- проблему відсутності гнучкого механізму щодо введення даних – за допомогою використання шаблонів.

Поданий алгоритм пристосований для генерації тексту англійською мовою. В подальшому для роботи з російськими та українськими словниками алгоритм потрібно модифікувати з точки зору врахування флективності цих мов.

Література

1. *Розенталь Д. Э. и др*. Словарь лингвистических терминов [Електронний ресурс http://www.gunner.info/bibliotek_Buks/Linguist/DicTermin/index.php] дата візиту 06.03.10.

2. Вікіпедія [Електронний ресурс http://ru.wikipedia.org/wiki/Семантическая_сеть] дата візиту 07.03.10.