УДК 004.912

М.М. Стовбчатий

ЗАСТОСУВАННЯ ГРАФОВИХ МОДЕЛЕЙ ТЕКСТУ ДЛЯ РОЗВ'ЯЗАННЯ ЗАДАЧ КОМП'ЮТЕРНОЇ ЛІНГВІСТИКИ

Вінницький національний технічний університет

Анотація

У даній статті розглянуто перспективи та методи аналізу текстової інформації, що забезпечують розпізнавання мовлення та визначення семантичних характеристик тексту. Основна увага приділена графовим моделям тексту. Наведено огляд основних можливостей лінгвістичного пакету NLTK – інструментального засобу, що пропонується для реалізації розглянутих моделей.

Ключові слова: аналіз; комп'ютерна лінгвістика; граф; модель; інформаційні технології; природна мова; розпізнавання мовлення; NLTK; парсер.

Annotation

This article examines the prospects and methods of analyzing text information that provide speech recognition and the definition of semantic characteristics of the text. The focus is on graph text patterns. The article provides an overview of the main features of the NLTK linguistic package - a tool that is proposed for the implementation of the models under consideration.

Key words: analysis; computer linguistics; earl; model; Information Technology; natural language; speech recognition; NLTK; parser

Дедалі більше людині доводиться взаємодіяти з комп'ютерними системами – як настільними, так і портативними та віддаленими. Технології розпізнавання мовлення дають змогу здійснювати цю взаємодію найбільш природним для людини чином, зокрема голосом [1].

Загалом, метою розпізнавання мовлення як розділу наукової дисципліни розпізнавання образів є отримання різного роду інформації на основі вхідного мовленнєвого (голосового) сигналу: про що говориться, хто говорить, якою мовою, в якому фізичному стані перебуває диктор тощо.

Ось доволі вичерпний перелік проблем, які вирішуються в ділянці розпізнавання та розуміння мовлення:

- автоматичне перетворення мовленнєвого сигналу на текст;
- введення інформації голосом, диктувальна машина;
- пошук ключових слів і фраз у потоці мовлення;
- смислова інтерпретація голосових повідомлень;
- ідентифікація та верифікація диктора;
- адаптація до голосу диктора та акустичного каналу;
- розпізнавання мови, якою говорить диктор, його акценту;
- усний переклад з однієї мови на іншу;
- розпізнавання емоційного та фізичного стану мовця.

Потрібно визнати, що найбільш складними задачами з цього переліку є пошук ключових слів і фраз у потоці мовлення, а також смислова інтерпретація голосових повідомлень. Мається на увазі, що розв'язання цих задач виконується в умовах якісно перетвореного мовленнєвого сигналу на текст, оскільки остання задача цілком досяжна з точки зору сучасних інформаційних технологій [1].

Отже, однією з найбільш вагомих проблем штучного інтелекту в цілому та комп'ютерної лінгвістики зокрема вважають визначення сенсу тексту. Для розв'язання цієї проблеми зазвичай використовують різні методи пошуку ключових слів у тексті, а також побудови онтологічної структури текстів, що дає можливість аналізувати головні поняття, про які йдеться у тексті на природній мові. Адекватною моделлю для представлення текстової інформації є граф, який визначається як кінцевий набір об'єктів (вершин) і множини пар різних вершин (рисунок 1.1). Така структура добре вивчена з точки зору математики і часто служить зручним засобом представлення структурованої інформації для подальшого аналізу. Графи використовуються в гуманітарних областях знань для автоматичної обробки текстів, інформаційного пошуку, реферування та індексування текстів, автоматичного перекладу, стилістичної діагностики, у завданнях атрибуції анонімних текстів тощо.

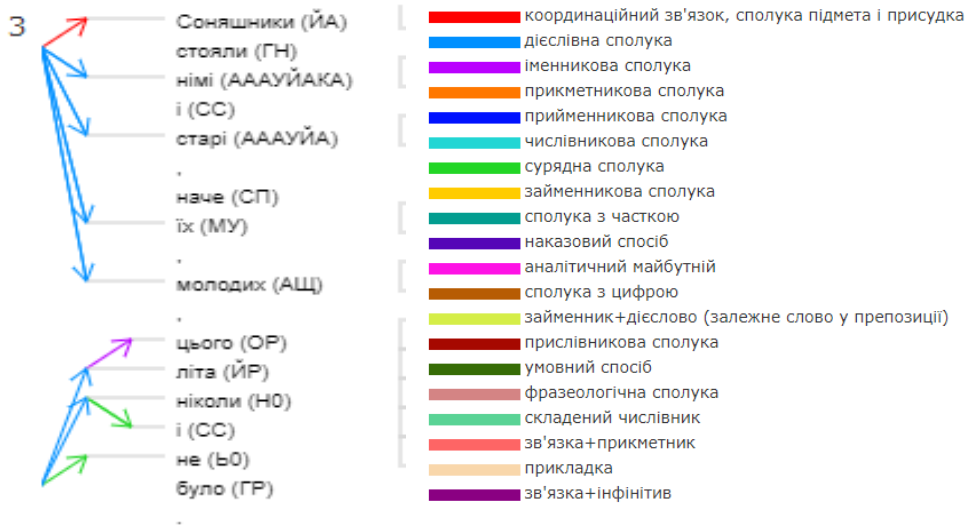

Рисунок 1.1 – Речення у вигляді графа

В основі моделі дерева лежить уявлення про побудову речення як про послідовне попарне синтагматичне зчеплення складових від мінімальних – окремих слів, до максимальної - речення, складовими якого, в загальному випадку, є група підмета і група присудка.

Подання синтаксичної структури в термінах дерева складових добре узгоджується з традиційним «розбором» речення, при якому підмет, присудок та інші елементи описуються категоріальним характеристиками – іменами частин мови або груп.

Граф – це засіб представлення змісту у вигляді семантичної зв'язку, це перекодування, структуризація знання. В основі перекодування лежить угруповання, смислова організація матеріалу. Граф виступає як засіб, з одного боку, фіксації структурно смислових компонентів тексту, а з іншого боку, контролю мовного продукту, тобто як механізм звірення і оцінки відповідності значення і форми деякої мовної структури до ідеалу [2].

У загальному вигляді методика побудови графа денотативної структури полягає в тому, що з тексту виділяються об'єкти (денотат), про які йдеться в тексті, і та інформація, яка повідомляється про них. Ця інформація виражається у вигляді інших об'єктів, пов'язаних з першими певними відносинами. Імена виділених об'єктів фіксуються у вигляді вершин графа, які розташовуються в певній послідовності. Ця послідовність визначається поданням про перехід від теми до підтемах, від загального до конкретного, від цілого до частини. При побудові такого графа аналіз тексту проводиться глобально, тобто окремі пропозиції не обов'язково виступають в якості одиниць аналізу. Тому розташування вершин графа може не відповідати порядку проходження об'єктів в тексті, а відображати скоріше їх об'єктивне, предметне співвідношення.

Всі операції, пов'язані з побудовою графа, носять змістовний характер. Побудова такого графа є своєрідний спосіб матеріалізації результату розуміння. Сама графічна форма робить структуру

відношень між елементами різних рівнів тексту наочною і доступною для огляду, сприяє більш повному і глибокому розумінню змісту тексту.

Як показала практика, подібна методика формалізації тексту дозволяє прискорити процес засвоєння, осмислення і репродукції тексту, допомагає учням, використовуючи різні мовні моделі, складати свої варіанти монологічних висловлювань по темі, досить швидко виводить їх у вільну комунікацію, сприяє формуванню семіотичної компетенції. Подібна кодова модель комунікації дозволяє підтримувати основоположні здатності природного інтелекту – здатність до абстрагування та згортки знання [2].

Для запропонованого в роботі [3] методу визначення ключових слів та схожих задач комп'ютерної лінгвістики, потрібно визначити зв'язки між словами в реченнях, що дасть змогу створити список пар зв'язаних слів тексту. Для цього потрібно зробити парсеринг (синтаксичний аналіз) кожного з речень, що дасть нам дерево залежностей слів у реченнях. З цією метою пропонується застосувати лінгвістичний пакет NLTK, якмй має інтерфейси для роботи з двома з найбільших і найпотужніших лінгвістичних парсерів: Stanford parser та MaltParser.

NLTK є однією з провідних платформ для побудови програм мовою Python для роботи з даними природної мови [4]. Даний програмний пакет забезпечує зручні у використанні інтерфейси для більш ніж 50 корпусів і лексичних ресурсів, таких як WordNet, одночасно з набором бібліотек обробки тексту для класифікації, токенізації, лемінізації та інших інструментів для роботи розв'язання задач комп'ютерної лінгвістики.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Тарануха В. Ю. Інтелектуальна обробка текстів Частина 1 / В.Ю. Тарануха. – К.: КНУ ім. Шевченка, 2014. – 235 с.

2. Севбо И. П. Графическое представление синтаксических структур и стилистическая диагностика / Ирина Платоновна Севбо. – Місто: Киев: Наукова думка, 1981. – 132 с.

3. Бісікало О. В. Метод визначення ключових слів англомовного тексту на основі DKPRO CORE / Бісікало О.В, Яхимович О.В. – "Технологічний аудит та резерви виробництва". – Том 1, № 2(21). – 2015. – С. 12 – 14.

4. Steven Bird Natural Language Processing with Python Analyzing Text with the Natural Language Toolkit / Steven B., Ewan K., Edward L // Sebastopol: O`REILLY. – 2010. – P. 504 – 512.

Стовбчатий Максим Михайлович, студент групи 2СІ-14б, факультет комп'ютерних систем та автоматики ВНТУ, м. Вінниця

Науковий керівник: Бісікало Олег Володимирович, д.т.н., професор, декан факультету комп'ютерних систем та автоматики, Вінницький національний технічний університет, м. Вінниця