УДК 004.8

Р.О. Петрук, аспір.,
Житомирский державний технологічний університет

ВИКОРИСТАННЯ ГРАФА СЕМАНТИЧНИХ ЗВ'ЯЗКІВ ДЛЯ ПОБУДОВИ МОДЕЛЕЙ ПРИРОДНИХ МОВ

(Представлено к.т.н., доц. Левицьким В.Г.)

Проаналізовано сучасні проблеми формалізації природних мов. Запропоновано використання графа семантичних зв'язків для описання структури природних мов та алгоритм побудови їх формальної моделі за допомогою аналізу даного графа.

Вступ. На даний час створення природномовних інтерфейсів пов'язане з двома основними проблемами. Перша – розпізнавання мови, тобто перетворення звукової інформації на текст. Друга – інтерпретація природномовних запитів або команд. Остання проблема є найменш дослідженою і найбільш актуальною. Їй приділяють велику увагу як виробники програмного та апаратного забезпечення комп'ютерних інформаційних систем, так і провідні наукові заклади. Найбільш важливі аспекти цієї проблематики вказані роботі [1].

У праці [2] показано, що у лінгвістичній алгебрі під ім'ям предиката слід розуміти не один, а цілу сім'ю пов'язаних один з одним предикатів. Це пов'язано з тим, що одна й та ж лексема у різних контекстах може мати абсолютно різні значення. У цьому випадку стандартні засоби для обробки формальних мов є безсилими. Для обробки природномовних текстів пропонуються складні механізми, робота яких іноді взагалі є прихованою від наочного спостереження. Декілька таких алгоритмів та моделей мови, побудованих у результаті їх роботи пропонується у роботах [3–5].

Всі алгоритми, призначені для побудови формальних моделей природних мов, базуються на аналізі текстів заданої мови, у тому числі й статистичному. Але, використовуючи ці методи, слід пам'ятати, що значний вплив на коректність їх роботи має зміст самих текстів. На сучасному етапі відкритими залишаються такі важливі проблеми, як аналіз семантичної подібності лексичних одиниць мови (синонімія, вживання різних слів у одному і тому ж значенні тощо), визначення місця у текстах, де одне і те ж слово залежно від сфери вживання має різне значення [6], аналіз некоректних за будовою текстів (неповні речення, стилістичні спотворення слів та словосполучень тощо), аналіз еквівалентів слова – словосполученя, які вживаються як єдина лексична одиниця [7].

В [8] описано принцип побудови моделі природної мови на основі графа семантичних зв'язків.

Основна частина досліджень. Для отримання більш повної моделі пропонується будувати граф семантичних зв'язків з метою його подальшого аналізу. Він являє собою орієнтований граф, вершини якого описують окремі лексеми, а ребра – зв'язки між лексемами.

Перед тим, як розглядати параметри компонент графа семантичних зв'язків, введемо поняття міжкомпонентної відстані для двох вершин.

Міжкомпонентна відстань – це різниця між порядковими номерами лексем у межах одного речення.

Параметром вершини v є кількість входжень відповідної лексеми $f(v)$ у текст відповідної мови. Параметром ребра e є вектор, перша компонента якого описує міжкомпонентну відстань для пари лексем $d(e)$, з'єднаних цим ребром, друга ж компонента описує кількість пар лексем із відповідною міжкомпонентною відстанню $f(e)$. Слід зазначити, що отриманий граф має особливість: дві вершини графа можуть бути з'єднані одночасно кількома ребрами, направленими однаково, але з різними параметрами $d(e)$.

Розглянемо приклад. Нехай мова містить три речення: „Лижники мають бажання купити лижі"; „Лижі можна купити у магазині"; „Лижники не мають бажання іти у магазин". Тоді граф буде мати вигляд, як зображено на рис. 1.

Одержаний граф описує мову з досить високою точністю. Для отримання загальної структури мови слід об'єднати лексеми за їх семантичною подібністю і описати зв'язки між отриманими групами. Для того, щоб говорити про семантичну подібність лексем, введемо поняття контексту.

Лівим контекстом $C_L(l)$ лексеми l називається множина пар $\langle e, v \rangle$, компонента e якої являє собою ребро, що входить у вершину графа, яка описує лексему l, а компонента v являє собою вершину, з якої виходить ребро e.

Правим контекстом $C_R(l)$ лексеми l називається множина пар $\langle e, v \rangle$, компонента e якої являє собою ребро, що виходить із вершини графа, яка описує лексему l, а компонента v являє собою вершину, з якої виходить ребро e.

Контекстом $C(l)$ лексеми l називається об'єднання лівого та правого її контекстів.

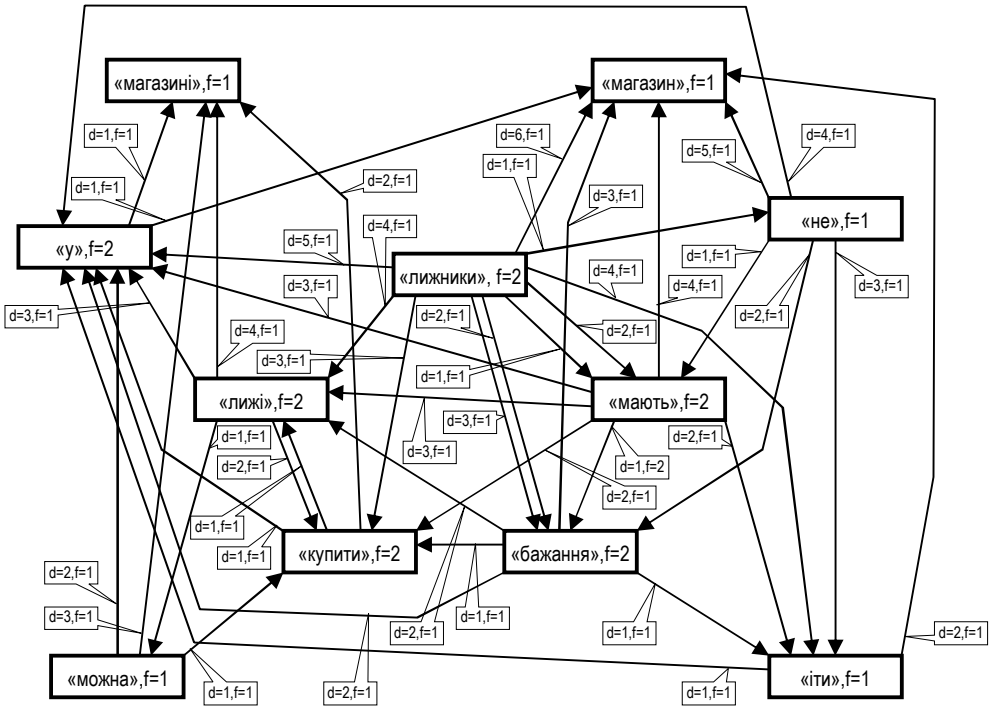

Рис. 1. Приклад графа семантичних зв'язків

Для прикладу розглянемо контекст слова „лижі" з попереднього прикладу (рис. 2). Глибина контексту дорівнює двом.

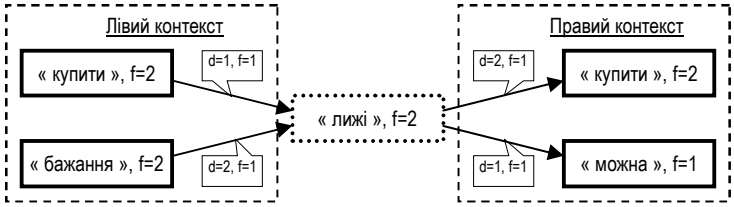

Рис. 2. Зразок контексту лексеми

У першому наближенні семантично подібними називають лексеми, у яких контексти є однаковими або мають незначну різницю.

Зрозуміло, що порівнювати попарно всі лексеми – завдання досить складне. Але слід звернути увагу на те, що семантично подібні лексеми мають знаходитись в межах одного і того ж контексту. Тому для прискорення пошуку семантично подібних лексем область перебору пропонується ділити на підобласті, кожна з яких є контекстом деякої лексеми.

У випадку виявлення пари подібних лексем вони об'єднуються в одну із дотриманням наступних правил.

1. Як нова лексема приймається множина, утворена шляхом об'єднання лексем із об'єднуваних вершин.

2. Кількість входжень отриманої лексеми дорівнює сумі кількостей входжень лексем до об'єднання.

3. Ребра графа, що входили в об'єднувані вершини, стають такими, що входять у новостворену вершину.

4. Ребра графа, що виходили з об'єднуваних вершин, стають такими, що виходять із новоствореної вершини.

5. Якщо у результаті об'єднання вершин з'являються ребра, що мають однакові початкову вершину, кінцеву вершину та міжкомпонентну відстань, то такі ребра об’єднуються в одне, причому кількість появи відповідних пар стає рівною сумі появ до об'єднання ребер.

Після об'єднання вершин пошук семантично подібних лексем починається спочатку. Алгоритм припиняє свою роботу тоді, коли не буде знайдено жодної пари семантично подібних лексем.

Висновки. Під час проведення досліджень було проаналізовано найпростіші частотні характеристики природних мов на прикладі російської. Показано можливість їх використання з метою визначення стилістичного забарвлення тексту. Ці характеристики можуть бути використані з метою відокремлення еквівалентів слова серед послідовностей окремих лексем мови.

Крім того, в даному дослідженні запропоновано схему алгоритму побудови формальної моделі природних мов за допомогою аналізу графа семантичних зв’язків.

Слід зауважити, що в подальшому необхідно більш детально проаналізувати частотні характеристики вживання у мові окремих лексем та їх комбінацій на предмет визначення важливих параметрів мови. Згідно з визначеними параметрами потрібно уточнити правила порівняння контекстів, а також, враховуючи ці зміни, потрібно уточнити (або, у разі потреби, навіть модифікувати) алгоритм побудови моделі мови.

ЛІТЕРАТУРА:

1. *Сулейманов Д.Ш.* Аналитический обзор отечественных и зарубежных работ в области обработки естественного языка в аспекте прагматически ориентированного подхода // Электрон. журнал Казанского госуниверситета „Информационные технологии”. – 1999.
2. *Шабанов-Кушнаренко Ю.П.* Теория интеллекта. Математические средства. Х.: Выща школа, 1984.
3. *Марченко О.О.* Алгоритми семантичного аналізу природномовних текстів: Дис. канд. фіз.-мат. наук. – К., 2005.
4. *Брусенцев В.А.* Алгебрологические модели формализации семантики предложений и их применение в информационных системах искусственного интеллекта: Дис. канд. техн. наук. – Х., 2003.
5. *Комісаренко Д.Ю.* Формалізоване проектування природномовних діалогових комп'ютерних систем: Дис. канд. техн. наук. – Вінниця, 2001.
6. *Струганець Л.В.* Динаміка лексичних норм в українській лексикографії ХХ століття: Дис. д-ра філолог. наук. – К., 2002.
7. *Лучик А.А.* Еквіваленти слова в українській і російській мовах: Дис. д-ра філолог. наук. – К., 2001.
8. *Петрук Р.О.* Побудова формальної моделі природної мови // XIII Всеукраїнська наукова конференція „Сучасні проблеми прикладної математики та інформатики”: Тези доповідей. – 2006. – С. 118.

ПЕТРУК Роман Олександрович – аспірант Житомирського державного технологічного університету.

Наукові інтереси:

– характеристики та моделі природних мов;
– використання платформи .NET для створення прикладних програмних продуктів.

E-mail: roman.petruk@rambler.ru

Подано 23.10.2006

Петрук Р.О. Використання графа семантичних зв’язків для побудови моделей природних мов

Петрук Р.А. Использование графа семантических связей для построения моделей естественных языков.

Petruk R.O. Using of graph of semantic relations for construction of models of natural language

УДК 004.8

Використання графа семантичних зв’язків для побудови моделей природних мов / Р.О. Петрук

Проаналізовано сучасні проблеми формалізації природних мов. Запропоновано використання графа семантичних зв'язків для описання структури природних мов та алгоритм побудови їх формальної моделі за допомогою аналізу даного графа.

УДК 004.8

Использование графа семантических связей для построения моделей естественных языков / Р.А. Петрук

Проанализированы современные проблемы формализации естественных языков. Предложено использование графа семантических связей для описания структуры естественных языков и алгоритм построения модели естественного языка с помощью анализа данного графа.

УДК 004.8

Using of graph of semantic relations for construction of models of natural language / R.O. Petruk

Modern problems of natural languages formalization are analyzed. Using of graph of semantic relations for description of structure of natural languages and algorithm for model of natural language construction with using of this graph are proposed.