

КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ  
ім. Тараса Шевченка  
ФАКУЛЬТЕТ КІБЕРНЕТИКИ

ЗВІТ  
про виконання лабораторної роботи  
з курсу «Штучний інтелект»  
тема: Система автоматичної індексації

Виконали: Процик Петро  
Криlach Олександр  
Група – ТТП-5  
Викладач: Марченко

КИЇВ 2005

# **ЗМІСТ**

1. Постановка задачі

2. Проектування та аналіз

3. Алгоритмічні особливості системи

4. Висновки

5. Література

## **Постановка Задачі**

Розробити систему семантичної індексації текстів. В якості допоміжних засобів можна використати семантичну базу слів англійської мови WordNet та систему синтаксичного аналізу для англійської мови - Grammar Link Parser (GLM).

## **Проектування та аналіз задачі**

Оскільки система розроблялась колективом авторів, були вирішено використовувати модульну архітектуру при проектуванні та реалізації системи семантичного індексування. Для того, щоб мати змогу використовувати засоби WordNet та GLM при розробці системи, необхідно розробити модулі та інтерфейси, які давали б можливість доступу до необхідної функціональності зазначених засобів. Було вирішено реалізувати такі інтерфейси у вигляді бібліотек динамічної компоновки (dll).

Крім того, система містить ще один зовнішній модуль, на який покладено задачі лексичного аналізу та виділення цілісних речень з природомовного тексту. Проблема виділення речень не є тривіальною задачею, як це часто може здаватись на перший погляд. Слід лише згадати проблеми пов'язані з пошуком кінця речення (оскільки символ «.» може виступати не лише як признак завершення речення), обробкою діалогу, прямого мовлення, нестандартних символів, тощо.

Внутрішня будова системи складається з трьох концептуально різних етапів обробки інформації. На першому, та самому «грубому» етапі, проводиться частотний аналіз вхідного тексту, тобто підраховується кількість входжень різних слів в текст. Для більш точних результатів в системі використовується морфологічний аналіз WordNet та список «стоп слів», тобто слів які не має необхідності обробляти. Вважається, що слова із «стоп списку» не несуть ні якої корисної інформації.

Результати цього етапу використовуються при виділенні речень, які в подальшому будуть підлягати семантичній обробці.

Виділення речень виконується на основі їх відносної ваги. Відносна вага речення визначається на основі частотного аналізу. Вага речення тим більше чим більше слів, які зустрічаються «часто», воно містить.

Далі відбувається синтаксичний аналіз виділених речень, тобто побудова дерев синтаксичного виведення. Ця задача покладена на модуль GLP. Результатом другого етапу є список дерев синтаксичного виведення.

На третьому етапі виконується аналіз списку дерев синтаксичного виведення з метою з'ясування до якого класу вони найбільш імовірно належать. Третій етап реалізований двома з багатьох можливих варіантів.

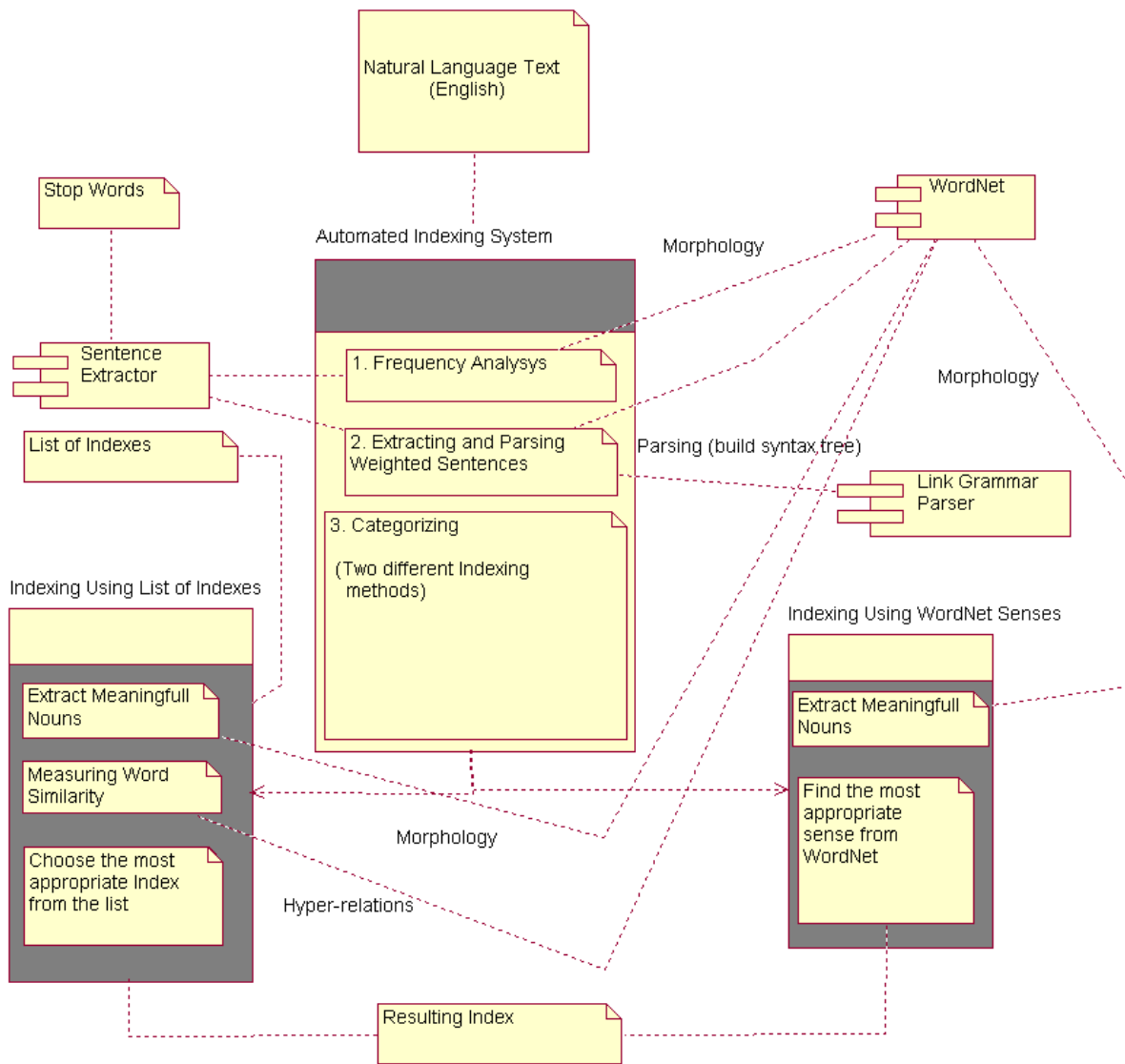
Перший варіант ґрунтується на використанні списку індексів. Кожне речення аналізується на степінь належності до кожного з індексів. Степінь належності визначається на основі результатів вимірювання семантичної близькості слів. В якості результату обирається той індекс до якого найбільш імовірно належать найбільше виділених речень.

Другий варіант ґрунтується на використанні семантичної мережі WordNet та відношення "is-a". При цьому в якості індексу тексту обирається той концепт з мережі який покриває найбільшу кількість найважливіших концептів з вихідного тексту. Важливими в нашому випадку будемо вважати концепти, які можна виділити з обраних та синтаксично проаналізованих на попередньому етапі речень.

Загальна структура та схема взаємодії модулів приведена на малюнку № 1 на наступній сторінці:

# Малюнок №1

(Структурна схема системи семантичного індексування)



## Алгоритмічні особливості системи

В цьому розділі будуть розглянуті особливості алгоритмічної реалізації деяких процедур, що використовуються у системі.

### Виділення речень з природомовного тексту.

Розглянемо деякі проблеми пов'язані з виділенням речень з природомовних текстів та їх подолання в системі.

Однією з проблем яка стоїть на заваді коректному виділенню речень є скорочення слів які завершуються символом «.», наприклад:

How old is **Mr.** John?

Найбільш коректним, на думку авторів системи, було б використовувати виділення речень разом з використанням системи синтаксичного аналізу текстів. Оскільки, в такому випадку подібні неоднозначні ситуації можна вирішувати шляхом використання відомостей про синтаксичну структуру речення. Проте такий метод потребує значно більших обчислювальних ресурсів та унеможливує пряме використання засобів LQP (так як він орієнтований на обробку цільних речень).

Слід зауважити, що для задач індексації можна обмежитись використанням списку слів-скорочень. Тобто, необхідно наперед задати слова-скорочення після яких може стояти крапка і в такому випадку її не слід сприймати як символ кінця речення.

Використання такої методики при достатньо повному списку слів-скорочень дає коректні результати на рівні 85% для текстів художньої літератури.

### Зваження речень

На другому етапі обробки виконується виділення речень на основі зваження. Це необхідно для коректного зваження об'ємів даних (тобто без значної втрати змісту) які будуть підлягати ресурсоємким процедурам обробки. Зваження речень – це один з підходів до зменшення оброблюваних даних. Для обчислення ваги речення використовується наступна формула:

$$Weight(S) = \frac{\sum_{word \in S} Max(Freq(word), 1)}{|S| * \underset{w \in AllWords}{Max}(Freq(W))}$$

,де

S – речення,

word  $\in$  S - слово яке міститься в реченні,

AllWords – всі слова в тексті

|S| - довжина речення

Freq(C) - частота зустрічання слова C в тексті

В якості іншого підходу можна запропонувати випадково-імовірнісний підхід. Тобто, обирати речення з тексту випадково, причому випадкова величина може мати статистично розраховану функцію розподілу, з метою виділення значимою інформації. Така функція розподілу може бути отримана на основі аналізу великого об'єму інформації певної предметної галузі та оцінки концентрації найважливіших концептів (їх фізичного розміщення в тексті). Такий підхід в реалізації не відображений.

### **Проблема неоднозначного синтаксичного розбору**

Деякі речення допускають неоднозначний синтаксичний розбір. В системі обирається довільний, оскільки проведені тести показали, що на предметну галузь тексту спосіб розбору речення впливає не суттєво.

### **Проблема виділення концептів з дерева синтаксичного аналізу**

Будемо вважати, що основний зміст тесту міститься в іменниках. Хоча необхідно зазначити, що в деяких випадках такий підхід може давати зовсім протилежні результати. Але ґрунтуючись на результати тестування та міркування мінімізації обчислювальних ресурсів такий підхід себе виправдовую.

Тому в системі з дерев синтаксичного аналізу виділяються іменники, що знаходяться в листях з коренем поміченим NP (Noun phrase). Результати виділення також залежать від результатів частотного аналізу

## Визначення подібності концептів

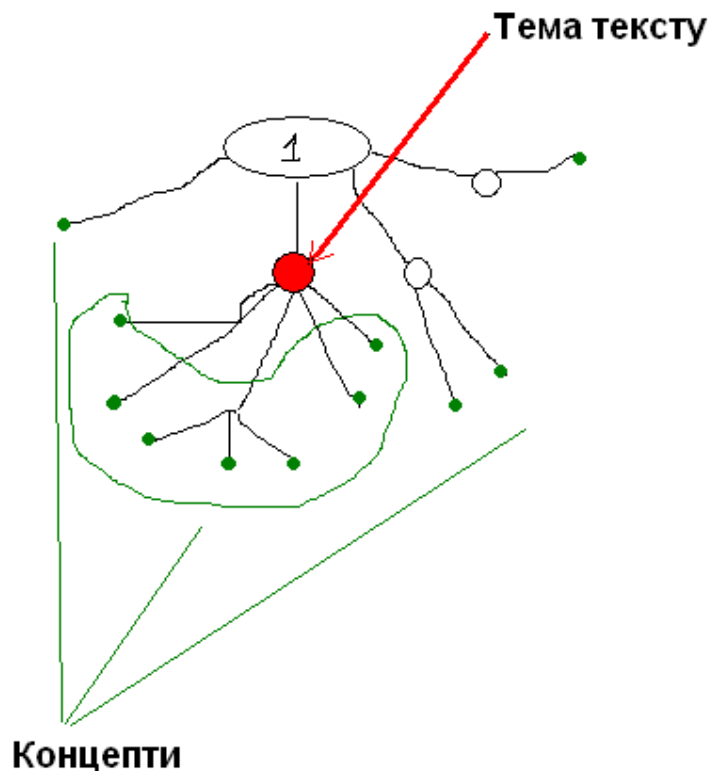
Розв'язання проблеми визначення подібності концептів необхідне в задачі класифікації речення.

Слово відноситься до того індексу, до якого воно найбільш семантично подібне. Речення відноситься до того індексу, до якого належать більшість його слів.

Семантична подібність концептів визначається на основі семантичної мережі WordNet. Алгоритм базується на знаходженні найкоротшої відстані між концептами на мережі (графі).

## Визначення найбільш загального концепту

Задача полягає у визначенні концепту, синами якого є більшість основних концептів тексту. Задачу можна проілюструвати такою схемою:



Для розв'язку цієї задачі використовується семантична мережа WordNet, а зокрема відношення Нурег, тобто «батько – син».

Наведемо фрагмент реалізації цього алгоритму з вихідного коду системи:

---

```
for (ttl = ws; ttl ; ttl=ttl->next)
{
  s1 = findtheinfo_ds(ttl->word, NOUN , -HYPERPTR, ALLSENSES);
  if(!s1)
    s1=findtheinfo_ds(morphstr(ttl->word,NOUN),NOUN,-HYPERPTR, ALLSENSES);
  if (s1)
    for (il = s1; il!=NULL; il = il -> nextss)
    {
      depth = 0;
      for (ssl = il; ssl!=NULL; ssl = ssl -> ptrlist) depth++;

      depth1 = 0;
      for (ssl = il; ssl!=NULL; ssl = ssl -> ptrlist)
      {
        depth1++;
        if (depth - depth1 <= low_sence_barier &&
            depth - depth1 >= high_sence_barier )
        {
          WIndex *w = add_to_index(ssl->words[0],ttl->word,Indexes);
          w->weight++;
        }
      }
    }
  free_syms(s1);
}
max = 0;
for (tw = *Indexes; tw; tw=tw->next)
  if (tw->weight > weight) { weight = tw->weight; max = tw;}
return max;
```

---

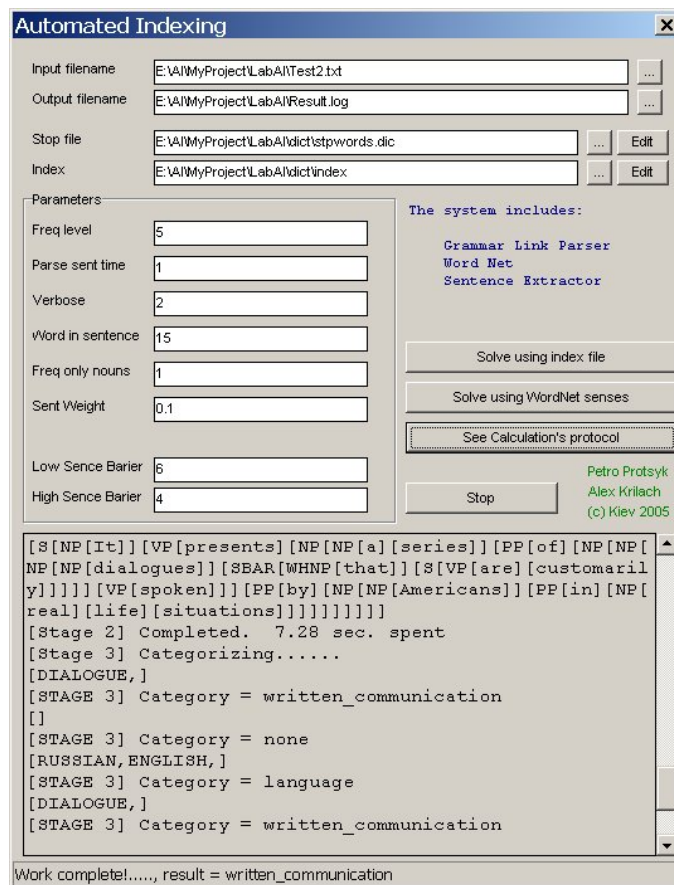
## Висновки

В результаті виконання лабораторної роботи нами була розроблена система автоматичного семантичного індексування. Розробка системи потребувала глибокого вивчення як технічних так і теоретичних питань пов'язаних з обробкою природомовних текстів. В якості допоміжних засобів в системі використані семантична мережа WordNet та синтаксичний аналізатор Link Grammar Parser.

Результати тестування системи показали цікаві результати пов'язані з різними аспектами англійської мови. Так було встановлено, що в детективних романах Артура Конан Дойля використано близько 5000 різних слів, причому 75% з них у тексті зустрічаються лише по одному разу.

На тестових прикладах система показала достатньо коректні результати, що дає нам змогу констатувати часткове розв'язання проблеми автоматичної ідексації.

На останок приведемо зовнішній вигляд ситеми з завантаженим прикладом:



## Література

- [1]. Alexander Budanitsky, Lexical Semantic Relatedness and Its Application in Natural Language Processing
- [2]. Daniel D.K. Sleator, Davy Temperley, Parsing English with a Link Grammar
- [3]. A. Budanitsky, G. Hirst, Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures
- [4]. W. Lewis, Measuring Conceptual Distance Using WordNet: The Design of a Metric for Measuring Semantic Similarity
- [5]. Inderjeet Mani, Automatic Summarization
- [6]. Э.В. Попов, Общение с ЭВМ на естественном языке, проблемы искусственного интеллекта, Москва, «Наука» 1982
- [7]. Victor Raskin, Semantic Ontology