computing@computingonline.net
www.computingonline.net

ISSN 1727-6209
International Journal of Computing

ФОРМУВАННЯ ЛІНГВІСТИЧНОЇ ОНТОЛОГІЇ НА БАЗІ СТРУКТУРОВАНОГО ЕЛЕКТРОННОГО ЕНЦИКЛОПЕДИЧНОГО РЕСУРСУ

Антон Михайлюк [1)], Олена Михайлюк [2)], Олексій Пилипчук [2)], Володимир Тарасенко [2)]

1) Київський університет імені Бориса Грінченка,Тимошенко, 13-Б, Київ, 04212, Україна, may-62@ukr.net
2) НТУУ «КПІ», кафедра СПСКС, просп. Перемоги, 37, Київ, 03056, Україна.
mes@scs.ntu-kpi.kiev.ua, ilexcorp@ukr.net, vtarasen@scs.ntu-kpi.kiev.ua

Резюме: *Вирішення задачі інтелектуалізації інформаційно-пошукової діяльності вимагає застосування допоміжних спеціалізованих лінгвістичних ресурсів. Одним із таких ресурсів може бути лінгвістична онтологія предметної галузі. Стаття розглядає підхід до організації програмних засобів для автоматизованого формування онтологічної бази знань шляхом конвертації структурованого енциклопедичного ресурсу у відповідні об'єкти онтології. Розглядаються процедури створення понятєвої бази онтології, ієрархії понять та мережі асоціативних зв'язків. Проводиться дослідження якісного та кількісного складу сформованої експериментальної онтології на основі українського сегменту Вікіпедії.*

Ключові слова: *лінгвістична онтологія, семантичні відношення, структурована енциклопедія.*

A CREATION OF THE LINGUISTIC ONTOLOGY BASED ON A STRUCTURED ELECTRONIC ENCYCLOPEDIC RESOURCE

Anton Mykhailiuk [1)], Olena Mykhailiuk [2)], Oleksiy Pylypchuk [2)], Volodymyr Tarasenko [2)]

[1)] Borys Grinchenko Kyiv University, 13-b Tymoshenko str., Kyiv, 04212, Ukraine, may-62@ukr.net
[2)] NTUU «KPI», SP SCS, 37 Peremogy avenue, Kyiv, 03062, Ukraine
mes@scs.ntu-kpi.kiev.ua, ilexcorp@ukr.net, vtarasen@scs.ntu-kpi.kiev.ua

Abstract: *Solving the problem of intelligent information retrieval requires the use of additional specialized linguistic resources. One of such type of resources is a linguistic ontology of a subject field. The paper considers an approach to develop software for an automatic creation of an ontological knowledge base by the converting structured encyclopedic resource into appropriate ontology's objects. The procedures of creation of the ontology entities base, the hierarchy of entities and the network of associative links are considered. Qualitative and quantitative analysis of the produced experimental ontology based on Ukrainian part of Wikipedia is investigated.*

Keywords: *linguistic ontology, semantic relations, structured encyclopedia.*

ВСТУП

Інтелектуалізація процедур аналізу контенту та інформаційно-пошукової діяльності вимагає розробки спеціалізованих лінгвістичних ресурсів, які б могли бути використані для підвищення ефективності інформаційно-аналітичних програмних засобів. Це дало б суттєвий поштовх розвитку таких лінгвістичних ресурсів, як словники синонімів, комп'ютерні тезауруси [1], семантичні мережі [2] тощо. Зокрема, останнім часом значно посилився інтерес до застосування онтологій під час роботи з текстомісткими об'єктами, що не випадково, оскільки онтологія за своєю природою має відображати структуру людських знань про оточуючий світ. Це дозволяє використовувати її для привнесення змістовної компоненти в процес обробки інформаційних об'єктів, їх аналізу, а

також в процедури інформаційного пошуку по цих об'єктах. Поняття онтології дуже широко трактується як в загально філософському сенсі, так і безпосередньо в рамках інформаційних технологій. Це пов'язано, в першу чергу, з тим, що онтології можуть різнитись за рівнем подання, за призначенням, за сферою застосування та методикою формування тощо. Оскільки однією з найпоширеніших галузей, де застосування онтології вбачається доцільним та перспективним, є сфера інформаційно-пошукової діяльності (в першу чергу це пошук [3, 4] або інформаційний моніторинг [5] текстових об'єктів в глобальній або локальній інформаційній мережі), тому онтологію доцільно розглядати з позиції лінгвістики, оскільки об'єкти мають текстову природу. Лінгвістична онтологія – це спеціальна база знань, що описує поняття зовнішнього світу та відношення між ними. Разом з тим це особливий клас онтологій, де поняття формуються на основі мовних одиниць, що відносяться до певної предметної галузі [1]. Можна виділити наступні важливі риси для структурно-логічних властивостей подібних онтологій:

- кожне поняття предметної галузі подається в онтології за допомогою синсета – сукупності близьких синонімів, що мають подібне значення;
- кожний синсет має певний зміст, що подається за допомогою унікального тлумачення;
- різні поняття можуть мати однакове мовне подання, в такому разі можна говорити про багатозначні терміни, а смислове розрізнення таких понять можливе за допомогою тлумачень;
- всі синсети об'єднуються в єдину ієрархічну таксономію понять – від абстрактних понять до конкретних;
- синсети онтології зв'язані між собою за допомогою семантичних відношень, найпоширеніші з яких – це відношення асоціації та гіпонім-гіперонім [1].

Крім того, використання онтологій в автоматичному режимі для синтетичних мов, до яких належить і українська мова, вимагає доповнення онтології нормальними формами слів для понять онтології.

Формування онтології програмними засобами передбачає аналіз певного текстового масиву даних, виокремлення понять та ідентифікацію зв'язків між ними у відповідності до заданих семантичних відношень. На жаль, така процедура вимагає надто складного апарату семантичного аналізу, який втім не гарантує якісної реалізації онтології через неоднозначність інтерпретації природомовних текстів. Вирішенням цього питання, на нашу думку, може бути залучення колекції документів енциклопедичного (словникового) характеру, що мають чітку структуру та створюються експертами. Саме такою структурованою колекцією для нас вбачається вільна електронна енциклопедія – Вікіпедія [6]. Її статті являють собою опис того чи іншого поняття, при цьому описова частина має цілком чітку структуру і, крім того, містить посилання на інші статті. Це робить можливим автоматизацію обробки таких документів та формування онтології.

Отже мета даної статті полягає в розробці способів організації програмних засобів, які б дозволили в автоматичному або напівавтоматичному режимі створити лінгвістичну онтологічну базу знань на основі статей певного мовного сегменту (напр., українського сегменту) Вікіпедії для застосування в інформаційно-пошуковій діяльності, зокрема в процедурах квазісемантичного пошуку [7].

1. СТРУКТУРА ВХІДНИХ ДАНИХ

Робота програмних засобів формування онтології має використовувати особливості структурної організації матеріалів Вікіпедії. Створення статей відбувається з використанням спеціального формату MediaWiki [8] (Рис.1).

```
<page>
  <title>Кортеж</title>
  <id>19488</id>
  <revision>
    <timestamp>2010-01-21T18:45:56Z</timestamp>
    <contributor>
      <username>Igor Yalovecky</username>
    </contributor>
    <text xml:space="preserve">'''Кортéж''' або "n"-ка  – в [[математика|математиці]] впорядкована та [[скінченна множина|скінченна]] сукупність елементів …
==Дивіться також==
* [[Декартів добуток]]
* [[Формальна мова]]
[[Категорія:Теорія множин]]
[[Категорія:Реляційна модель даних]]
[[en:Tuple]] [[ru:Кортеж]]
    <text>
  </revision>
</page>
```

Рис.1 – Фрагмент XML-документа статті Вікіпедії

Цей формат дозволяє однозначно описувати структурні елементи статті (заголовки, розділи, посилання, мета-елементи тощо), що в свою чергу дає змогу ефективно в автоматичному режимі їх обробляти. Весь архів колекції певного мовного сегменту статей Вікіпедії зберігається у вигляді XML-документа і доступний для завантаження. XML-розмітка дає змогу виділяти необхідні вузли для подальшої обробки. На рис.1 наведений приклад вузла, що описує одну із статей із українського сегмента. Даний фрагмент ілюструє структуру XML-вузла для статті, що описує поняття «Кортеж». Як видно із цього фрагмента, даний вузол містить не лише вихідний код статті, але й деяку допоміжну метаінформацію. Власне для аналізу контенту інтерес становлять два поля – це поле <title> (заголовок статті) та поле <text> (безпосередньо текст статті).

Усю множину статей Вікіпедії за різними критеріями можна віднести до декількох груп, що зведені в табл. 1.

Аналіз структури Вікіпедії дозволяє зробити висновок про те, що статті Вікіпедії разом із взаємними внутрішніми посиланнями створюють певний прототип онтологічної бази знань [6]. Основну роль тут відіграють тлумачні статті та статті категорій, а група статей багатозначних понять виконує допоміжну роль при обробці тлумачних статей по кожному зі значень. Крім того, незавершені статті можуть бути використанні для ідентифікації зв'язку між іншими статтями, якщо ті будуть містити на неї посилання.

Таблиця 1. Види статей Вікіпедії

Вид статей	Опис
Тлумачні статті	Статті описують певне поняття, подію або явище. Відповідно до назви є основним джерелом інформаційного наповнення енциклопедії та відповідно служать головним джерелом для створення онтології.
Статті, що описують багатозначне поняття	Цей вид статей призначений для зберігання списку усіх наявних на поточний момент у відповідному сегменті Вікіпедії значень деякого терміну. Такі статті містять посилання на відповідну тлумачну статтю, якщо така є, і короткий унікальний опис по кожному з семантичних значень терміна. Основний індикатор таких статей – це наявність службової мітки *{{disambig}}* на початку текстової частини.
Статті категорій	Статті, що описують поняття-категорію із загальної ієрархії категорій Вікіпедії. В тлумачних статтях одне або декілька таких понять можуть вказуватись як батьківські категорії.
Статті, що описують файли	Файлові статті описують файлові об'єкти Вікіпедії (наприклад, зображення) та містять специфічну для відповідних об'єктів інформацію – посилання, розмір, тип тощо.
Незавершені статті	В тлумачних статтях часто зустрічаються посилання на статті, які з тих чи інших причин ще не написані. Вони не мають описової частини, але мають заголовок.
Службові статті	Такі статті не мають безпосереднього інформаційного навантаження, спрямованого безпосередньо на користувача, і використовуються в процесі розробки Вікіпедії. Зокрема це можуть бути шаблони статей, довідки, зауваження, обговорення і т. ін.

2. СТРУКТУРА ВИХІДНИХ ДАНИХ

Виявлені особливості структури Вікіпедії дозволяють розглядати програмне забезпечення для формування лінгвістичної онтології як засіб своєрідної конвертації елементів Вікіпедії у об'єкти онтології. Оскільки останні будуть активно використовуватись в інформаційно-пошуковій діяльності, питання зберігання таких об'єктів з можливістю ефективного та оперативного доступу до них може бути вирішено шляхом застосування системи баз даних. Відповідно до логічної організації онтології, на рис. 2 запропонована структурна схема реляційної бази даних, яка містить основні компоненти онтології та зв'язки між ними. Результатом формування інформаційного наповнення онтології будуть заповнені відповідними значеннями таблиці бази даних. Розглянемо детальніше всі реляційні відношення в наведеній базі даних.

Synset – центральне реляційне відношення, яке відображає синсет онтології. Воно містить наступні атрибути: унікальний ідентифікатор (id), символьне подання синсета українською (ua), російською (ru) та англійською (en) мовами, унікальне смислове тлумачення синсету (descr).

Поля ru та en введені для встановлення потенційного зв'язку україномовної онтології з аналогічними онтологіями, створеними на основі російського та англійського сегментів Вікіпедії. В залежності від потреб, множину мов можна як розширювати, додаючи відповідні поля до складу реляційного відношення синсету, так і зовсім відмовитись від додаткових мовних атрибутів, залишивши лише назву синсету основною мовою.

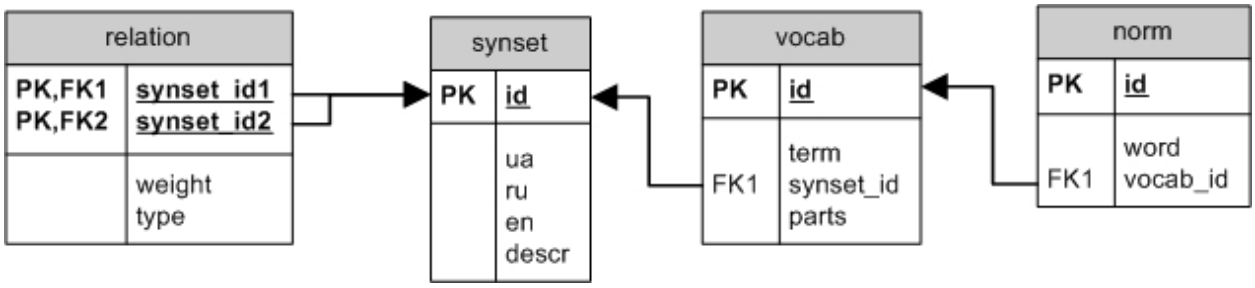

Рис. 2 – Структура бази даних онтології

Vocab – допоміжне реляційне відношення, призначене для зберігання усіх синонімічних входів синсету. Воно складається з унікального ідентифікатора (id), терміну – символьного подання синонімічного входу (term), ідентифікатора батьківського синсету (synset_id) та поля, що містить кількість слів присутніх в терміні (parts).

Norm – допоміжне реляційне відношення для зберігання нормальних форм слів, що входять до складу символьного подання терміну з Vocab. Складається з унікального ідентифікатора (id), символьного подання нормальної форми слова з терміну (word) та ідентифікатора батьківського терміну з Vocab (vocab_id).

Relation – ключове реляційне відношення, що відображає наявні зв'язки між поняттями онтології. Поля synset_id1 та synset_id2 містять ідентифікатори синсетів, між якими присутній семантичний зв'язок. Поле weight характеризує цей зв'язок певним значенням ваги, чим більше це значення, тим сильнішим вважається зв'язок між синсетами. Нарешті, поле type зберігає тип семантичного відношення, так в контексті квазісемантичного пошуку розглядаються два типи відношень – асоціація та гіпонім-гіперонім.

3. АЛГОРИТМІЧНА ОРГАНІЗАЦІЯ ФОРМУВАННЯ ОНТОЛОГІЇ

Процес формування онтології можна розділити на декілька етапів: підготовчий етап, етап створення бази синсетів, етапи побудови ієрархічних та асоціативних зв'язків, а також заключного етапу корекції. Наведений на діаграмі (рис.3) клас mediawiki_parser забезпечує виконання всіх вказаних кроків по створенню онтології. Оскільки вихідні дані статей Вікіпедії подаються в форматі XML-документа, необхідний початковий етап, основна мета якого – це парсинг XML даних та підготовка їх до подальшої обробки. Згідно фрагмента таких даних, що були представлені вище, в процесі створення онтології необхідно послідовно переходити до кожної наступної статті (вузол page). Із піддерева вузла page для створення онтології далі становлять інтерес заголовок (вузол title) та текст статті (вузол text).

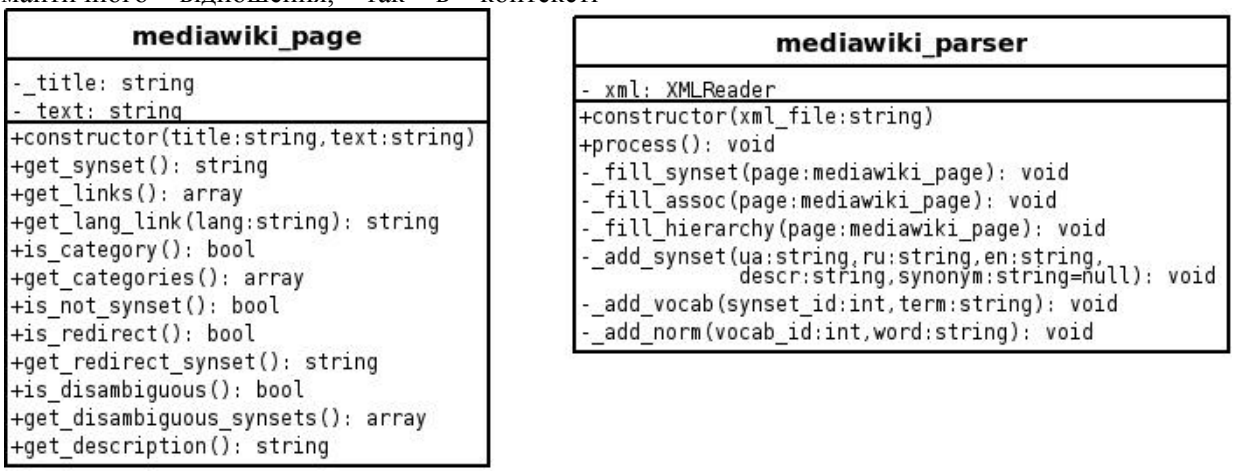

Рис. 3 – Діаграма класів для роботи з Вікіпедією у форматі MediaWiki.

На основі отриманих даних створюється екземпляр класу mediawiki_page для роботи зі сторінкою Вікіпедії у форматі MediaWiki (рис. 3). Він дозволяє оперувати даними статті, робити різні перевірки, виділяти логічні фрагменти тощо.

Відповідно до діаграми (рис. 3) клас mediawiki_page реалізує наступні важливі методи:

- get_synset – повертає символьне подання синсету для даного типу статті (якщо це можливо);
- get_links – аналізує текст статті та повертає масив посилань на інші статті Вікіпедії;
- get_lang_link(lang) – повертає посилання на відповідну статтю, написану іншою мовою відповідно до параметра lang, якщо така існує;
- is_categoгy – перевіряє, чи має поняття, що описується даною статтею, статус категорії;
- get_categories – повертає перелік категорій, до яких належить дана стаття;
- is_not_synset – перевіряє, чи може поняття, що описується даною статтею, інтерпретуватися як синсет;
- is_redirect – перевіряє, чи є дана стаття лише перенаправленням на іншу статтю;
- get_redirect_synset – повертає символьне подання синсету, на статтю якого здійснюється перенаправленням з даної статті;
- is_disambiguous – перевіряє, чи не є дана стаття описом багатозначного поняття;
- get_disambiguous_synsets – повертає перелік синсетів з однаковим написанням, але з різними тлумаченнями, сформований відповідно до вмісту статті багатозначного поняття;
- get_description – повертає тлумачення для поняття, що описується даною статтею.

Всі наведені вище методи класу активно використовуються на відповідних етапах формування онтології. Після підготовчого етапу настає черга заповнення безпосередньо бази даних онтології. Сам процес інкапсульований у методі process класу mediawiki_parser. На рис.4 наведений псевдокод цього методу.

```
/* Створюємо XMLReader */
xml = new XMLReader(xml_file);
/*Формуємо базу синсетів*/
/* Поки можливо, зчитуємо наступний вузол page */
while (xml_page = xml->next('page')) do
        /*Створюємо об'єкт статті*/
        page = new mediawiki_page(xml_page->get('title'), xml_page->get('text'));
        /*Запускаємо метод заповнення синсетів на основі поточної статті*/
        this->_fill_synset(page);
end while;
/*Повертаємо вказівник XMLReader на початок файлу*/
xml->reset();
/*Формуємо зв'язки онтології*/
/* Поки можливо, зчитуємо наступний вузол page */
while (xml_page = xml->next('page')) do
        /*Створюємо об'єкт статті*/
        page = new mediawiki_page(xml_page->get('title'), xml_page->get('text'));
        /*Запускаємо метод заповнення ієрархічних зв'язків*/
        this->_fill_hierarchy(page);
        /*Запускаємо метод заповнення асоціативних зв'язків*/
        this->_fill_assoc(page);
end while;
```

Рис.4. Псевдокод заповнення бази даних онтології

На етапі створення бази синсетів фактично формується основа онтології, її понятєва складова. В кожній статті шляхом аналізу її заголовка і тексту визначається тип поняття, яке вона описує, і відповідним чином заповнюються таблиці synset, vocab та norm. Для цього використовуються методи класу mediawiki_parser: _add_synset, _add_vocab та _add_norm. Необхідно зазначити, що метод _add_synset автоматично додає інформацію в словник синсетів за допомогою виклику _add_vocab, який в свою чергу автоматично викликає метод _add_norm. На рис.5 наведений псевдокод, який абстрактно показує послідовність дій на даному етапі (метод _fill_synset класу mediawiki_parser).

Етап створення ієрархічних зв'язків онтології, як зазначалось раніше, базується на використанні ієрархії статей Вікіпедії, що описують категорії. В кінці кожної тлумачної

статті чи статті, що описує категорію, вказується перелік батьківських категорій, до яких така стаття належить. Неважко помітити, що в рамках Вікіпедії одне поняття може відноситись одразу до декількох категорій. Ієрархічна структура, сформована таким чином, буде відрізнятись від класичної ієрархії понять, де в кожного поняття є лише одне більш загальне поняття (інакше кажучи, лише один батьківський елемент). Однак таке обмеження не відповідає реальним відношенням між поняттями, оскільки для більшості із них неможливо однозначно виявити тільки один батьківський елемент. Саме тому пряма конвертація всіх категоріальних зв'язків Вікіпедії в ієрархічну структуру синсетів лінгвістичної онтології розглядається як найоптимальніша, за умови, що при проектуванні інструментарію, який буде використовувати безпосередньо гіперонімічні зв'язки онтології буде врахована вказана вище особливість формування таких зв'язків. Відповідно до цього псевдокод методу _fill_hierarchy буде виглядати як на рис.6.

Етап створення асоціативних зв'язків принципово відрізняється від етапу створення ієрархічних зв'язків лише тим, що тут для ідентифікації зв'язків використовуються посилання на інші статті Вікіпедії (а отже й інші поняття в онтології) в тексті поточної статті. Крім того, оскільки посилання може зустрічатись неодноразово, а також можуть мати місце зворотні зв'язки (коли стаття, на яку здійснюється перехід, містить зворотні посилання), це дає змогу підвищувати вагу даного відношення на деяку величину *delta*, тим самим відображаючи посилення семантичного відношення. Чим більше таких посилань, тим сильніший зв'язок між статтями, відповідно тим більшою має бути вага відношення. Для того, щоб розподілення ваги зв'язків по онтології було достатньо рівномірним, величину *delta* варто вибирати порівняно невелику по відношенню до початкового значення ваги w_0. Заповнення асоціативних зв'язків відбувається в методі _fill_assoc, псевдокод якого подано на рис.7.

```
/*Якщо стаття описує не синсет*/
if page->is_not_synset() then
        return; /*припиняємо роботу*/
/*Якщо стаття описує перенаправлення на іншу статтю*/
else if page->is_redirect() then
        /*Додаємо новий синсет з символьним поданням поняття із статті,*/
        /* на яку йде перенаправлення, в якості синоніма подаємо поняття поточної статті*/
        this->_add_synset(page->get_redirect_synset(), '', '', '', page->get_synset());
/*Якщо стаття описує багатозначне поняття*/
else if page->is_disambiguous() then
        /*Отримуємо перелік всіх синсетів та їх тлумачень*/
        list(synset,description) = page->get_disambiguous_synsets();
        /*Для кожного запису із переліку створюємо новий синсет*/
        for each record in list(synset,description) do
                /*Додаємо новий синсет synset з тлумаченням description*/
                /*в якості синоніма передаємо поточний заголовок статті*/
                this->_add_synset(synset, '', '', description, page->get_synset());
        end for;
/*Якщо стаття описує поняття або категорію*/
else
        /*Додаємо новий синсет*/
        this->_add_synset(
                page->get_synset(),
                page->get_lang_link('ru'),
                page->get_lang_link('en'),
                page->get_description());
end if;
```

Рис.5. Псевдокод заповнення бази синсетів онтології

```
/*Якщо стаття не є описом поняття або категорії*/
if page->is_not_synset() OR  page->is_redirect() OR  page->is_disambiguous() then
        return; /*припиняємо роботу*/
/*Якщо стаття описує поняття або категорію*/
else
        /*Для кожної категорії статті*/
        for each category in page->get_categories() do
                /*Зберігаємо відношення ієрархії з вагою відношення рівне w0*/
                database->relation->insert(
                        'synset_id1' = category,
                        'synset_id2' = page->get_synset(),
                        'type' = 'H',
                        'weight' =w0
                );
        end for;
end if;
```

Рис.6. Псевдокод формування ієрархічних зв'язків онтології

```
/*Якщо стаття не є описом поняття або категорії*/
if page->is_not_synset() OR  page->is_redirect() OR  page->is_disambiguous() then
        return; /*припиняємо роботу*/
/*Якщо стаття описує поняття або категорію*/
else
        /*Для кожного посилання на іншу статтю*/
        for each link in page->get_links() do
                /*Якщо дане відношення уже було занесено в базу даних*/
                if database->relation->select_where(
                        'synset_id1' == page->get_synset(),
                        'synset_id2' == link) then
                        /*Оновлюємо вагу відношення на деяку величину delta*/
                        line = database->relation->select_where(
                                'synset_id1' == page->get_synset(),
                                'synset_id2' == link);
                        line->update(weight = weight + delta);
                else
                        /*Зберігаємо відношення асоціації з вагою відношення рівне w0*/
                        database->relation->insert(
                                'synset_id1' = link,
                                'synset_id2' = page->get_synset(),
                                'type' = 'A',
                                'weight' = w0
                        );
                end if;
        end for;
end if;
```

Рис.7. Псевдокод формування асоціативних зв'язків онтології

Нарешті, останній етап – це необов'язковий етап корекції уже створеної на попередніх стадіях онтологічної бази. Він може відбуватися як автоматично (наприклад, видалення можливих службових понять та всіх їх зв'язків), так і шляхом «ручного» редагування онтологічної бази редакторами онтології (наприклад, видалення нерелевантних або помилкових зв'язків тощо). Така корекція може знадобитись зважаючи на неповну прозорість автоматичного формування онтології та наявність службової інформації в першоджерелі, яка може відображатись у специфічних для енциклопедії статтях (напр., наявність категорії «статті на букву А» тощо), включати які в онтологію недоцільно.

4. РЕЗУЛЬТАТИ ТА ВИСНОВКИ

Для створення і аналізу експериментальної лінгвістичної онтології розроблено дослідний прототип програмного забезпечення конвертації

українського сегменту Вікіпедії у відповідну україномовну онтологічну базу знань. Результати роботи програмного модуля формування лінгвістичної онтології на основі статей українського сегменту Вікіпедії на момент проведення експерименту зведено у табл. 2. Наведені тут кількісні характеристики дають підстави стверджувати, що така онтологія охоплює досить широку область людського знання і може бути використана в роботі інформаційних систем. Необхідно зробити наголос на тому, що створена таким чином онтологія – це лише початковий етап, базис для подальшої роботи експертів і в той же час науково-інформаційний ресурс для дослідницько-експериментальної роботи в області пошукових технологій. В процесі подальшої роботи над онтологією можна вносити нові поняття, привносити нові зв'язки чи видаляти помилкові.

Табл.2. Кількісні характеристики онтології

Характеристика	Кількісний показник
Кількість синсетів	243948
Середня кількість слів у синсеті	2.24
Кількість ієрархічних зв'язків	127366
Кількість асоціативних зв'язків	2658584

Якісний аналіз створеної онтології в цілому провести фактично неможливо, зважаючи на дуже велику кількість понять та зв'язків між ними. Цей процес вимагає довготривалого використання такої онтології на практиці та потребує значних людських та часових ресурсів. Підвищення якості онтологічної бази знань має відбуватись паралельно із її використанням в розробці процедур інформаційного пошуку, аналізу контенту тощо. Втім оціночну характеристику можна визначити шляхом вибіркового аналізу. Для цього будують декілька невеликих точкових підмереж (наприклад, таку, як зображено на рис.8) з різних частин онтології та оцінюють основні зв'язки між синсетами.

Розгляд таких фрагментів дозволяє зробити висновок, що поняття та відношення між ними, за деякими незначними винятками, в основному відповідають понятєвій структурі предметних галузей. Тому якісний склад онтології та адекватність відношень можна вважати цілком прийнятними для використання в інформаційно-пошуковій діяльності.

Рис. 8. Фрагмент асоціативних зв'язків сформованої онтології

Таким чином, наведені процедури формування онтології дають можливість порівняно легко в автоматичному режимі створити достатньо об'ємну лінгвістичну онтологію, що буде відповідати основним вимогам до таких ресурсів. Крім того, оскільки Вікіпедія, як основне джерело інформації для формування онтології, має властивість багатомовності, в перспективі можливе створення кросмовних онтологій, що значно розширить сферу їх застосування.

5. СПИСОК ЛІТЕРАТУРИ

[1] B.V. Dobrov, N.V. Lukashevich, The linguistic ontology of nature sciences and technologies for information retrieving applications, S*cientists' notes of Kazan. univ.,* 2 (2007), pp. 49-72. (in Russian)

[2] E.F. Skorohodko, *Semantic Networks and Automatic Text Processing*, Naukova Dumka, Kyiv, 1983, p. 217. (in Russian)

[3] D.V. Lande, *Knowledge Retrieving at the Internet. professional work*, Williams, Moscow, 2005, p. 272. (in Russian)

[4] D.V. Lande, A.A. Snarskiy, I.V. Bezsydnov, *Internetics. Navigation in the Complex Networks: Models and Algorithms*, Librikom, Moscow, 2009, p. 264. (in Russian)

[5] D.V. Lande, *The Base of Information Flows Integration*, Engeeniring, Kyiv, 2006, p. 240. (in Russian)

[6] Z.S. Syed, T. Finin, A. Joshi, Wikipedia as an ontology for describing documents, *In Proceedings of ICWSM*, AAAI Press, 2008, pp. 136-144.

[7] A.Y. Mykhailiuk, O.V. Pylypchuk, M.V. Snizhko, V.P. Tarasenko, Quasi-semantic search of textual data in an electronic

information source, *Radioelectronics and Informatics*, KhNURE, Kharkov, 3 (2009), pp. 61-67. (in Ukrainian)

[8] D.J. Barrett, *MediaWiki*, O'Reilly Media, 2008, p. 384.

***Антон Михайлюк**, доцент кафедри Інформатики Київського університету імені Бориса Грінченка. Кандидат технічних наук, старший науковий співробітник. Наукові інтереси: методи та засоби інтелектуального аналізу природномовних текстових інформаційних об'єктів, інноваційні підходи до комп'ютеризації науково-освітньої діяльності.*

***Олена Михайлюк**, науковий співробітник кафедри системного програмування і спеціалізованих комп'ютерних систем Національного технічного університету України «Київський політехнічний інститут». Закінчила Київський політехнічний інститут у 1989р. Наукові інтереси: експертні системи, їх застосування в задачах аналізу даних.*

***Олексій Пилипчук**, асистент кафедри системного програмування і спеціалізованих комп'ютерних систем Національного технічного університету України «Київський політехнічний інститут». Закінчив НТУУ «Київський політехнічний інститут» у 2008р.*

Наукові інтереси: інформаційний пошук, інформаційний-моніторинг, квазісемантичний пошук текстових даних, системи транспортної логістики.

***Володимир Тарасенко**, завідувач кафедри системного програмування і спеціалізованих комп'ютерних систем Національного технічного університету України «Київський політехнічний інститут». Доктор технічних наук, професор, заслужений діяч науки і техніки України, лауреат Державної премії України в галузі науки і техніки. Наукові інтереси: методи та засоби підвищення ефективності обробки ресурсів глобального електронного інформаційного простору.*

computing@computingonline.net
www.computingonline.net

ISSN 1727-6209
International Journal of Computing

A CREATION OF THE LINGUISTIC ONTOLOGY BASED ON A STRUCTURED ELECTRONIC ENCYCLOPEDIC RESOURCE

Anton Mykhailiuk [1)], Olena Mykhailiuk [2)], Oleksiy Pylypchuk [2)], Volodymyr Tarasenko [2)]

[1)] Borys Grinchenko Kyiv University, 13-b Tymoshenko str., Kyiv, 04212, Ukraine, may-62@ukr.net
[2)] NTUU «KPI», SP SCS, 37 Peremogy avenue, Kyiv, 03062, Ukraine
mes@scs.ntu-kpi.kiev.ua, ilexcorp@ukr.net, vtarasen@scs.ntu-kpi.kiev.ua

Abstract: *Solving the problem of intelligent information retrieval requires the use of additional specialized linguistic resources. One of such type of resources is a linguistic ontology of a subject field. The paper considers an approach to develop software for an automatic creation of an ontological knowledge base by the converting structured encyclopedic resource into appropriate ontology's objects. The procedures of creation of the ontology entities base, the hierarchy of entities and the network of associative links are considered. Qualitative and quantitative analysis of the produced experimental ontology based on Ukrainian part of Wikipedia is investigated.*

Keywords: *linguistic ontology, semantic relations, structured encyclopedia.*

1. INTRODUCTION

Intellectualization of content analysis procedures and information retrieval requires the development of specialized linguistic resources such as dictionaries of synonyms, thesauri computer [1], semantic networks [2], and, in particular, linguistic ontologies. So the purpose of this paper is to develop ways of organizing software that would allow to create linguistic ontological knowledge base, based on articles of a language segment (eg, the Ukrainian segment) of Wikipedia, in automatic or semi-automatic mode for using in retrieval activities, including procedures quasisemantic search [7].

2. INPUT DATA STRUCTURE

The work of software tools of ontology building use the structural organization features of Wikipedia materials. Special format MediaWiki is used in articles creation [8]. The whole archive collection of a language segment of Wikipedia articles stored as an XML-document and it is available for download. XML-layout allows to distinguish the necessary components for further processing. Fig. 1 shows the example of a site that describes one of the articles of the Ukrainian segment. The entire set of Wikipedia articles on various criteria can be attributed to several groups, which represented in Table 1. Analysis of the Wikipedia structure allows to do the conclusion that Wikipedia articles with mutual internal links create the prototype of an ontological knowledge base [6].The main role belongs here explanatory articles and categories articles; a group of articles of multivalued concepts has a supporting role in the processing of explanatory articles on each of the values. In addition, incomplete articles can be used to identify the relationship between other articles if they contain link to it.

3. OUTPUT DATA STRUCTURE

Detected features of the Wikipedia structure allows to consider software to form a linguistic ontology as a means of converting Wikipedia elements to ontology objects. Since they will be actively used in information-retrieval activities, that issues of storage such objects with the possibility of efficient and operative access to them can be solved by means of a database system application. According to logic of the ontology, the block diagram of a relational database with basic components of ontology and relations between them is proposed in Fig. 2. Database tables with filled corresponding values are the result of information filling of the ontology. Consider more detail all relations in the relational database. Synset – a central relation that reflects an ontology synset. It contains the following attributes: a unique identifier (id), symbolic representation of a synset by Ukrainian (ua), Russian (ru) and English (en) languages, unique semantic interpretation of the synset (descr).

The fields ru and en were introduced to establish a potential connection of the Ukrainian ontology with similar ontologies created on the basis of Russian and English segments of Wikipedia. Vocab – an auxiliary relation designed to store all synonymous inputs of the synset. It consists of a unique identifier (id), a term – a symbolic representation of the synonymic input (term), an identifier of a parent synset (synset_id) and a field that contains the number of words in the term (parts). Norm – an auxiliary relation for storage of normal forms of words which comprise to the symbolic representation of the term from Vocab. Relation – a key relation that reflects the existing connections between the ontology concepts. The fields synset_id1 and synset_id2 contain synset identifiers between which there is a semantic relationship. A field weight characterizes this relationship by certain value of the weight. The more the value, the stronger is the relationship between synsets. Finally, a field type stores the type of the semantic relation, so in the context of the quasisemantic search two types of relationships are considered – association and hyponym/hypernym.

4. ALGORITHMIC ORGANISATION OF ONTOLOGY FORMING

The formation of the ontology can be divided into several stages: the preparatory stage, the stage of creation of a synsets database, the stages of building of hierarchical and associative relationships and the final stage of correction. Mediawiki_parser class ensures the execution all these steps to create the ontology. According to the diagram (Fig. 3) mediawiki_page class implements the following important methods:

- get_synset – returns the symbolic representation of the synset for this type of an article (if it possible);
- get_links – analyzes the text of the article and returns an array of links to other Wikipedia articles;
- get_lang_link (lang) – returns a reference to the relevant article written in another language in accordance with the lang parameter, if it is;
- is_category – checks whether the concept described in this article has the category status;
- get_categories – returns a list of categories, which related with this article;
- is_not_synset – checks whether the concepts described in this article, can be interpreted as a synset;
- is_redirect – checks whether a given article only redirect to another page;
- get_redirect_synset – returns the symbolic representation of the synset, of which article the redirection from this article is executed;
- is_disambiguous – checks whether this article describes a multi-valued concept;
- get_disambiguous_synsets – returns a synset list with the same spelling but with different interpretations, formed in accordance with the content of a multivalued concept article;
- get_description – returns the interpretation of a concept, described in this article.

All specified class methods are widely used at the appropriate stages of ontology forming. After the preparatory phase the ontology database is filled. The process encapsulated in the process method of mediawiki_parser class. The basis of an ontology (its conceptual component) is formed actually at the stage of the synsets database creation. For each article the type of a concept is determined by analyzing its title and text, and synset, vocab and norm tables are filled. Mediawiki_parser class methods are used: _add_synset, _add_vocab and add norm. It should be noted that _add_synset method automatically adds information to the synsets dictionary by calling _add_vocab, which in turn automatically invokes the _add_norm method.The stage of creating of hierarchical relationships for ontology, as noted earlier, is based on the using of Wikipedia articles hierarchy that describes the categories. At the end of each explanatory article or article, which describes the category, the list of parent categories to which this article refers is specified. Links to other Wikipedia articles (and hence other concepts in the ontology) in the text of the current article are used for relationships identification in the stage of creating associative connections. It is the fundamental difference from the stage of creating of hierarchical relationships. In addition, because the reference can occur repeatedly, and there may be feedbacks, it allows to increase the weight of the relationship at some value delta. The associative relationships are filled in the _fill_assoc method.

5. RESULTS AND CONCLUSIONS

The research prototype of conversion software of Wikipedia's Ukrainian segment to the appropriate Ukrainian-ontological knowledge base for creation and analyzing of experimental linguistic ontology is developed. The work results of the software module of forming the linguistic ontology of Ukrainian segment of Wikipedia articles at the time of the experiment are shown in the Table 2.

These quantitative characteristics give reason to believe that such ontology covers quite a wide area of human knowledge and can be used in the work of information systems.

Table 2. Ontology quantitative characteristics

Characteristics	Quantitative parameter
Quantity of synsets	243948
Average number of words in a synset	2.24
Quantity of hierarchical relationships	127366
Quantity of associative relationships	2658584

6. REFERENCES

[1] B.V. Dobrov, N.V. Lukashevich, The linguistic ontology of nature sciences and technologies for information retrieving applications, S*cientists' notes of Kazan. univ.*, 2 (2007), pp. 49-72. (in Russian)

[2] E.F. Skorohodko, *Semantic Networks and Automatic Text Processing*, Naukova Dumka, Kyiv, 1983, p. 217. (in Russian)

[3] D.V. Lande, *Knowledge Retrieving at the Internet. professional work*, Williams, Moscow, 2005, p. 272. (in Russian)

[4] D.V. Lande, A.A. Snarskiy, I.V. Bezsydnov, *Internetics. Navigation in the Complex Networks: Models and Algorithms*, Librikom, Moscow, 2009, p. 264. (in Russian)

[5] D.V. Lande, *The Base of Information Flows Integration*, Engeeniring, Kyiv, 2006, p. 240. (in Russian)

[6] Z.S. Syed, T. Finin, A. Joshi, Wikipedia as an ontology for describing documents, *In Proceedings of ICWSM*, AAAI Press, 2008, pp. 136-144.

[7] A.Y. Mykhailiuk, O.V. Pylypchuk, M.V. Snizhko, V.P. Tarasenko, Quasi-semantic search of textual data in an electronic information source, *Radioelectronics and Informatics*, KhNURE, Kharkov, 3 (2009), pp. 61-67. (in Ukrainian)

[8] D.J. Barrett, *MediaWiki*, O'Reilly Media, 2008, p. 384.