УДК 004.421

СЕЛЕКТИВНИЙ ПІДХІД ДО АВТОМАТИЗОВАНОГО ФОРМУВАННЯ ОПЕРАТИВНОГО СПЕЦІАЛІЗОВАНОГО ЛЕКСИКО-СЕМАНТИЧНОГО СЛОВНИКОВОГО РЕСУРСУ

Т.М. Заболотня

НТУУ «КПІ», м. Київ, Україна, tatiana104@yandex.ru

Запропоновано підхід до програмного формування лексико-семантичного словникового ресурсу з певної предметної галузі на основі упорядкованої селекції частини вмісту універсального словника формату WordNet. Проаналізовано узагальнені особливості термінології документації різних сфер людської діяльності та, зокрема, сфери права. Розроблено послідовність відбору синсетів до спеціалізованого словника. Досліджено ефективність використання сформованого словника (у сенсі швидкодії відповідного програмного забезпечення та адекватності результатів його роботи) на прикладі задачі виправлення орфографічних помилок у предметно-орієнтованих текстових даних.

ЛІНГВІСТИЧНЕ ПРОГРАМНЕ ЗАБЕЗПЕЧЕННЯ, СПЕЦІАЛІЗОВАНІ МАШИННІ СЛОВНИКИ

Вступ

На сьогоднішній день для організації ефективної комп'ютерної обробки спеціалізованих текстових даних особливої значущості набуває врахування програмним забезпеченням особливостей термінології відповідної предметної галузі. На жаль, існуючі інформаційні системи працюють із тематичними електронними документами здебільшого на рівні їх формальних характеристик, а для більш детального аналізу текстів, як правило, застосовуються універсальні програмні засоби та звичайні словникові ресурси.

Між тим від коректності укладання комп'ютерних словників та алгоритмів обробки їх вмісту значною мірою залежить швидкість роботи відповідного програмного забезпечення із текстовими даними та адекватність результатів, які при цьому отримує користувач. Звідси, формування ефективного предметно-орієнтованого словникового ресурсу є однією з основних проблем побудови спеціалізованого програмного інструментарію аналізу текстових даних, вирішенню якої і присвячена дана стаття.

1. Постановка задачі

Зазвичай для лексикографічного опису терміносистеми предметної галузі будується машинний словник тезаурусного типу [1, 2, 3], який дозволяє вирішити проблему синонімії термінів та термінологічних сполучень, що зустрічаються в текстах. Чим більше понять та їх синонімів систематизовано у тезаурусі і чим різноманітніші види зв'язків між ними, тим ширші можливості інформаційної системи, яка його використовує. Але при цьому слід уникати введення надмірної кількості синонімів та рідковживаних слів, неістотних для текстів даної тематики.

На жаль розробка нового тезаурусу є доволі трудомістким завданням, а аналіз його вмісту не забезпечує повною мірою рішення задач автоматизованої обробки спеціалізованого тексту.

По-перше, відбір ключових слів з текстового матеріалу, виявлення умовної еквівалентності між ними, відбір дескрипторів, встановлення парадигматичних відношень між термінами і дескрипторами — все це є питаннями, вирішення яких потребує спільної роботи спеціалістів конкретної предметної галузі та філологів [3, 4]. Часто залучення таких професіоналів до роботи над створенням програмного тезаурусу неможливе. Тому актуальною є розробка методів швидкої та максимально автоматизованої побудови словника, який би підтримував виконання необхідного для lingware такого класу мінімуму операцій та який міг би виступити основою для подальшої «ручної» корекції терміносистеми відповідними фахівцями.

По-друге, комп'ютеризоване укладання спеціалізованого словника наново потребує врахування відповідним програмним забезпеченням специфіки лексики та характерних рис документів обраної предметної галузі. Це спричиняє перетворення таких програмних засобів на складні інтелектуальні системи, розробка яких не є швидшою та простішою за «ручне» створення тематичного словника. Вирішенням даної проблеми, на думку автора, є автоматизоване формування машинних словників на основі існуючих, коректно укладених професіоналами універсальних ресурсів шляхом відбору лексики з останніх за певними правилами. Результатом проведення модифікації наповнення словника є підвищення оперативності роботи програми, яка його використовує, за рахунок уникнення нею зайвої обробки слів, які взагалі не можуть бути присутні у текстах даної тематики.

По-третє, залучення тезаурусу для розв'язання питань, пов'язаних із обробкою окремих слів (наприклад, для виправлення орфографічних помилок), не є доцільним через те, що у ньому зберігаються поняття, котрі, як правило, виражені словосполученнями та цілими зворотами. З огляду

на це необхідним є укладання поряд із тезаурусами спеціалізованих лексико-семантичних словників, структурною одиницею яких є окрема лексема. Це дозволить розширити функціональність відповідних інформаційних систем (ІС) та зробить останні більш гнучкими у роботі.

Виходячи із наведених вище аргументів, **метою дослідження** стали розробка, теоретичне обґрунтування та експериментальна апробація підходу до формування оперативного предметно-орієнтованого словникового ресурсу шляхом автоматизованої модифікації існуючого універсального лексико-семантичного словника з урахуванням особливостей терміносистеми конкретної предметної галузі.

У відповідності до поставленої мети **задачами дослідження** є:

– визначення послідовності формування спеціалізованого словника шляхом упорядкованої фільтрації наповнення універсального лексико-семантичного словникового ресурсу;

– експериментальне дослідження ефективності використання словника з певної предметної галузі, отриманого відповідно до запропонованого підходу.

Для теоретичного обґрунтування та експериментальної апробації даного підходу було обрано одну з найбільш специфічних та формалізованих сфер людської діяльності, ефективна робота у якій безпосередньо залежить від швидкості та точності функціонування програмного інструментарію аналізу текстових даних, — *сферу права*.

2. Основні етапи автоматизованого формування спеціалізованого словникового ресурсу

2.1. Вивчення вимог до мови документів предметної галузі та їх впливу на вміст словника

Мова будь-якої предметної галузі характеризується цілою низкою особливостей, які повинні бути врахованими програмним забезпеченням обробки текстів. Тому для створення адекватного вузькоспрямованого словникового ресурсу необхідно проаналізувати існуючі вимоги до спеціалізованої мови, складені відповідними фахівцями, та визначити, які зміни в універсальному словнику може спричинити їх дотримання.

Отже, за характером впливу на наповнення словника вимоги до мови будь-якої предметної галузі (у т.ч. і до мови права) можна розділити на 2 групи [3, 5]:

1) вимоги, що визначають набір слів, які можуть застосовуватися у текстах даної предметної галузі, і набір слів, поява яких у документах виключена;

2) вимоги, які сприяють збереженню упорядкованості терміносистеми.

Так, наприклад, на основі результатів вивчення робіт спеціалістів у сфері юридичної та комп'ютерної лінгвістики автором було визначено найважливіші для розробки lingware вимоги до мови правових документів.

До першої групи можна віднести 2 узагальнюючі вимоги: **точність і ясність**. Серед вимог, які базуються на зазначених вище, найбільш поширеними є: емоційна /експресивна/ нейтральність лексики; простота та надійність мови права, що виключає багатозначність; економічність; високий ступінь абстрактності [2, 6].

З огляду на те, що система термінів кожної галузі науки (юридична терміносистема тут не є винятком) становить певну множину взаємопов'язаних елементів, які створюють чітку єдність і цілісність, наділену інтегральними властивостями і закономірностями [2, 5], до вимог, які запобігають утворенню у словнику хаотичного набору слів, автором віднесено вимогу **системності**.

Для того, щоб задовольнити наведеним вимогам, формування спеціалізованого словника має проходити у два етапи:

– фільтрації лексики універсального словникового ресурсу (виключення з його складу слів, нехарактерних для документів даної предметної галузі);

– систематизації частини лексем, яка була відібрана на етапі фільтрації.

2.2. Фільтрація вмісту універсального лексико-семантичного словника

Для модифікації та налаштування на задану предметну галузь було обрано популярний лексико-семантичний словник формату WordNet, розроблений ученими Прінстонського університету [7] як такий, що вільно розповсюджується, широко використовується та легко локалізується. Базовою структурною одиницею WordNet є синонімічний ряд (синсет), що об'єднує слова із подібним значенням. На множині лексем W, які зберігаються у даному словнику, можна визначити функцію *getsynset*, яка дозволяє для кожного слова отримати множину відповідних йому синонімічних рядів з множини усіх синсетів словника S:

$$\begin{aligned} &\text{getsynset} : W \to S \\ &W \ni w \to \{s \mid s \in S\} \equiv \text{getsynset}(w). \end{aligned} \qquad (1)$$

Синсети в WordNet зв'язані між собою семантичними відношеннями, серед яких особливу роль грають гіпернімія та гіпонімія, які дозволяють організовувати синсети в ієрархічні структури (дерева) [8]. Приклад структури словника формату WordNet наведено на рис. 1.

Фільтрацію наповнення wordnet-словника пропонується здійснювати шляхом статистичної обробки репрезентативної вибірки текстів з масиву документів, який повинен аналізуватися за допомогою даного словника.

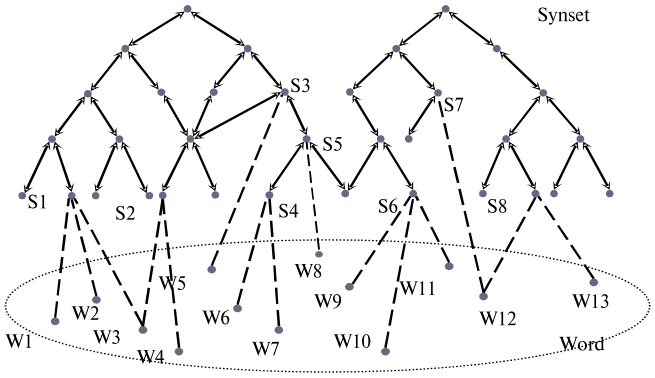

Рис. 1. Структурна організація словника формату WordNet

Крок 2.2.1. Вибір з документів усіх слів, їх лематизація та визначення абсолютної частоти появи відповідних лексем у тексті (на рис.1 таким словам відповідають вершини *w*1-*w*3, *w*5, *w*6, *w*12). Позначимо отриману множину слів W', де $W' \subset W$ (W — множина усіх лексем, що містяться у словнику). Вважаємо, що якщо слово з тексту не міститься у словнику, то воно помилкове, або є таким, що його можна не включати до подальшої обробки.

Крок 2.2.2. Знаходження синсетів, до складу яких входять відібрані на попередньому кроці слова (на рис.1 це відповідно синсети *s*1-*s*4, *s*7, *s*8). В результаті отримуємо множину синсетів слів S', де $S' \subset S$ (S — множина усіх синсетів, що містяться у словнику), таку що:

$$\forall s_t \in S' \quad \exists w_t \in W' : \text{getsynset}(w_t) = \{s \mid s \in S'\} \ni s_t \tag{2}$$

До обробки на наступному етапі потраплять лише слова та синсети, які належать множинам W' та S' відповідно (тобто частота появи яких у текстах документів більша, ніж нуль).

Якщо тексти документів, на основі аналізу яких було проведено фільтрацію вмісту словника, відповідають усім вимогам фахівців до мови предметної галузі, то у результаті здійснення описаних кроків мають бути отримані множини слів та відповідних їм синсетів, які так само задовольняють зазначеним вимогам [6]. Але, як можна бачити на рис. 1, вибрані таким чином синсети є розрізненими та несистематизованими. Цей недолік буде усунений на наступному етапі формування спеціалізованого словника.

2.3. Розширення множин W' та S' з метою виконання умови системності лексики

Крок 2.3.1. Побудова шляхів від кожного $s \in S'$ до найближчих за структурою словника синсетів.

Метою даного кроку є знаходження зв'язку між відібраними на попередньому етапі синсетами і, завдяки цьому, відновлення цілісності словника. Відстань *maxdist* (кількість дуг у шляху), в межах якої синсети вважаються близькими, визначається емпіричним шляхом.

Для виконання даного кроку на множині синсетів S визначається функція *shortpath* (див. (3)), за допомогою якої для кожного $s \in S'$ проводиться пошук шляхів до найближчих за структурою словника синсетів, які також належать множині S'. Результатом її виконання є масив, який складається з множин s_{path}, що містять синсети одного шляху між s_b та s_e. Синонімічні ряди, які входять до складу знайдених шляхів, заносяться до множини синсетів S_1''.

$$S_1'' = \{ s_{path} \subset S \mid s_{path} = \text{shortpath}(s_b, s_e)\}, \quad \text{де } s_b, s_e \in S' \wedge |s_{path}| < \text{maxdist}. \tag{3}$$

Якщо встановити *maxdist* = 5, то для прикладу, наведеного на рис.1, будуть побудовані шляхи між вершинами *s*2-*s*3, *s*2-*s*4 та *s*3-*s*4. Слід відзначити, що із подальшим збільшенням даної відстані потрібно поводитись обережно, тому що транзитивність на тезаурусних відношеннях працює тільки на «коротких» дистанціях [8].

Отже, даного кроку недостатньо для того, щоб отримати систематизований та цілісний підсловник з універсального wordnet-словника. Тому розширення множини синсетів на цьому не закінчується.

Крок 2.3.2. Побудова шляхів від кожного із синсетів $s \in S'$ до синсетів верхнього рівня.

На відміну від попереднього кроку, у даному випадку для формування шляхів використовується значно звужений набір відношень (це відношення типу «частина-ціле» та родо-видові, які дозволяють організовувати синсети у дерева [8]). За рахунок урахування даних ланцюжків синсетів, по-перше, відновлюється ієрархічність структури словника, а по-друге, синсети, між якими зв'язок ще не встановлено, можуть бути пов'язані через спільну вершину одного з верхніх рівнів. Синонімічні ряди, які входять до складу знайдених s_{path}, заносяться до множини синсетів S_2''.

Аналогічно до функції *shortpath*, визначимо на множині синсетів S функцію *toppath*:

$$S_2'' = \{ s_{path} \subset S \mid s_{path} = \text{toppath}(s_b, s_{top}) \}, \quad 45\ s_b \in S' \wedge s_{top} \in S. \tag{4}$$

У результаті виконання зазначених кроків щодо розширення множини синсетів S' отримуємо множину $S'' = S_1'' \cup S_2''$, якою і буде доповнений новий спеціалізований словник.

$$S_{new} = S' \cup S''. \tag{5}$$

де S_{new} — множина синсетів отриманого словника.

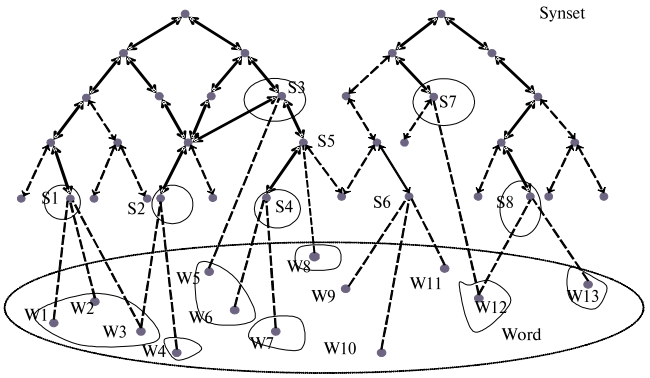

Рис. 2. Результат виконання етапу розширення множини синсетів wordnet-словника

Для прикладу, наведеного на рис. 2, множину S_{new} складають усі синсети, з'єднані жирною суцільною лінією.

Крок 2.3.3. Відображення отриманої множини синсетів S'' на множину W.

У межах даного кроку відбувається визначення множини слів, які входять до синсетів, відібраних раніше для розширення словника. Множину слів W'' можна визначити як:

$$W'' = \{w \in W \mid \text{getsynset}(w) \cap S'' \neq \varnothing\}. \quad (6)$$

У такому разі набір слів, які відповідають множині синсетів S_{new}, буде отримано за формулою (7):

$$W_{new} = W' \cup W''. \quad (7)$$

На рис. 2 множину W_{new} складають усі вершини w_i окрім $w9$, $w10$, $w11$.

Отже, результатом проведення фільтрації та наступної систематизації вмісту словникового ресурсу, мають бути відібрані множини синсетів та відповідних їм лексем, які складатимуть спеціалізований словник з певної предметної галузі (у нашому випадку — з права).

Щоб перевірити цілісність отриманого набору синсетів, використаємо **метод Мальгранжа-Томеску** для визначення компонентів сильної зв'язності в орграфах [9]. Процес модифікації словника пропонується вважати успішним, якщо до компонента сильної зв'язності, який містить найбільшу кількість синсетів, входять більше 90% вибраних синсетів.

Синсети, котрі належать множині S', але не увійшли до компонента сильної зв'язності, можна додати до неї «вручну», або збільшити значення *maxdist* та повторити для цих синсетів дії III етапу (див. вище). Наприклад, якщо для випадку, який розглядається у даній статті (див. рис. 2), збільшити *maxdist* до 6, між синсетами $s4$ та $s7$ можна побудувати шлях, а це означає, що відібраний підграф буде сильно зв'язним.

2.4. Додавання нових синсетів до сформованого спеціалізованого словника

Результатом виконання описаних вище дій є новий спеціалізований словник, побудований на основі універсального лексико-семантичного ресурсу. У межах даного етапу розглянемо, яким чином можна уточнювати наповнення вже існуючого словника. Слід зазначити, що механізм доповнення словника багато у чому подібний до вже розглянутих етапів, тому нижче виокремимо лише ті кроки, які найбільш відрізняються від попередніх.

Крок 2.4.1. Відбір з нових документів слів та відповідних синсетів, які ще не містяться у спеціалізованому словнику:

$$\begin{aligned} &W_{add} \subset (W \setminus W_{dict}),\ S_{add} \subset (S \setminus S_{dict}),\\ &\forall s_t \in S_{add}\\ &\exists w_t \in W_{add} : \text{getsynset}(w_t) \cap (S \setminus S_{dict}) = X \ni s_t,\\ &\text{де } X \neq \varnothing \end{aligned} \quad (8)$$

де W_{add}, S_{add} — множини відібраних слів та синсетів; W_{dict}, S_{dict} — множини слів та синсетів, які входять до спеціалізованого словника.

Звідси можна зробити висновок про те, що універсальний лексико-семантичний словник залучається до процесу уточнення спеціалізованого словникового ресурсу так само, як і на стадії створення останнього.

Крок 2.4.2. Побудова шляхів від кожного $s \in S_{add}$ до найближчих за структурою словника синсетів $s_x \in S_{add}$ та $s_y \in S_{dict}$.

Дії, виконання яких передбачається на даному кроці, дозволяють:

1) встановити семантичні зв'язки між вже сформованим спеціалізованим словником та множиною синсетів, відібраних на попередньому кроці:

$$\begin{aligned} &S_1 = \{ s_{path} \subset S \mid s_{path} = \text{shortpath}(s_b, s_e) \},\\ &\text{де } s_b \in S_{add} \wedge s_e \in S_{dict} \wedge |s_{path}| < \max\text{dist};\\ &S_2 = \{ s_{path} \subset S \mid s_{path} = \text{shortpath}(s_b, s_e) \},\\ &\text{де } s_b \in S_{dict} \wedge s_e \in S_{add} \wedge |s_{path}| < \max\text{dist}; \end{aligned} \quad (9)$$

2) встановити семантичні зв'язки всередині множини синсетів S_{add}:

$$\begin{aligned} &S_3 = \{s_{path} \subset S \mid s_{path} = \text{shortpath}(s_b, s_e)\},\\ &\text{де } s_b, s_e \in S_{add} \wedge |s_{path}| < \max\text{dist}. \end{aligned} \quad (10)$$

Множину синсетів, отриманих на даному кроці, можна визначити як:

$$S_{add}' = (S_1 \cup S_2 \cup S_3) \cap (S \setminus S_{dict}). \quad (11)$$

Крок 2.4.3. Побудова шляхів від кожного із синсетів $s \in S_{add}$ до синсетів верхнього рівня.

Здійснюється аналогічно до розглянутого вище відповідного кроку етапу розширення множини синсетів з метою дотримання умови системності лексики словника (див. крок 2.3.2).

Нехай результатом виконання даного кроку є множина S_{add}''. Тоді до словника будуть додані синсети $S_{add}' \cup S_{add}''$.

Відображення отриманої множини синсетів на множину W та перевірка цілісності допрацьованого спеціалізованого словника проводиться так само, як це було описано раніше.

Зазначимо, що повністю видаляти слова та синсети з універсального wordnet-словника, які не увійшли до спеціалізованого ресурсу, не доцільно, тому що у такому випадку розробники будуть позбавлені можливості вносити будь-які зміни до останнього. Тому пропонується ввести ранжування вмісту словника за ознакою частоти входження лексем та відповідних синсетів до вихідних текстів, а також до складу допоміжних побудованих ланцюжків в універсальному словнику. До обробки програмними засобами, які використовують модифікований словник, у такому випадку будуть залучатися слова, які мають показник частоти більший встановленого порогового значення.

3. Дослідження ефективності використання спеціалізованого словника

Для підтримки впровадження вищезазначених змін у wordnet-словник було створено допоміжне програмне забезпечення. Аналіз ефективності використання спеціалізованого словника проводився на прикладі задачі виправлення орфографічних помилок у вузькотематичних текстових даних: було порівняно швидкість роботи інформаційної системи з підключеним до неї універсальним і модифікованим словником та точність отриманих результатів. На основі одержаних даних можна зробити висновок, що побудований згідно з запропонованим підходом словник може виконувати усі функції універсального словника, але при цьому час обробки нашого словника та кількість невірних варіантів виправлення, які формуються програмою, відчутно зменшилися (для проведеного дослідження час обробки зменшився до 1, 5 разів, а кількість гіпотез виправлення — до 3 разів).

Висновки

Запропонований підхід до формування спеціалізованого лексико-семантичного словника передбачає його автоматизоване створення на основі існуючого універсального wordnet-ресурсу. Врахування особливостей терміносистеми конкретної предметної галузі при цьому дозволяє відбирати для нового словника релевантні лексеми та синонімічні ряди. Описані узагальнені етапи та окремі кроки щодо формування достатньо наповненого, цілісного та ієрархічно упорядкованого словника, за допомогою якого можна адекватно аналізувати текстові дані з певної тематики.

Дослідження ефективності використання створеного словникового ресурсу показали підвищення швидкодії та точності роботи програмного забезпечення, яке його використовує. Перевагами даного підходу також є зниження трудомісткості та часу розробки словника.

Таким чином, побудова лексико-семантичного словника з урахуванням особливостей мови конкретної сфери людської діяльності відіграє важливу роль при розробці спеціалізованого програмного забезпечення обробки текстових даних. Отримані внаслідок застосування запропонованого підходу результати можуть бути використані для подальшого вивчення проблеми створення програмного інструментарію лінгвістичної підтримки роботи спеціалістів будь-якої галузі.

Список літератури: 1. *Когаловский М.Р.* Перспективные технологи информационных систем. — М.: АйТи: ДМК, 2003. — 284 с. 2. *Язык закона* / Под ред. А.С.Пиголкина. — М., 1990 — 189 с. 3. *Артикуца Н.В.* Мова права і юридична термінологія [Навч.посіб. для студ. юрид. спец.вищ. навч. закл.]/ Нац. ун-т «Києво-Могилян. акад.», Центр інновац. методик правн. освіти. — [2-е вид., змін. і допов.]. — К.: Стилос, 2004. — 275 с. 4. *Карпіловська Є.А.* Вступ до комп'ютерної лінгвістики. — Донецьк: ТОВ «Юго-Восток, Лтд», 2003. — 184 с. 5. *Толста С.А.* Актуальні питання уніфікації та стандартизації юридичних термінів // Гуманітарна освіта в технічних вищих навчальних закладах. — 2003. — Вип. 7. — С. 64–69. 6. *Право і лінгвістика:* Матеріали II міжнар. наук.-практ. конф.: У 2-х ч. — Сімф. : ДОЛЯ, 2004. 7. *Wordnet* — a Lexical Database for English. Princeton University, Princeton, NJ, 2001 8.. *Марченко О.О.* Алгоритми семантичного аналізу природномовних текстів: Дис...канд.фіз.-мат. наук:01.05.01/ КНУ ім. Тараса Шевченка. — К., 2005. — 150 с. 9. *Новиков Ф.А.* Дискретная математика для программистов. — СПб: Питер, 2000. — 304 с.: ил.

Надійшла до редколегії 27.09.2007