УДК 811.161.2

А. Романюк, І. Сундутова, М. Романишин
Національний університет “Львівська політехніка”,
кафедра систем автоматизованого проектування,
кафедра прикладної лінгвістики

МЕТОДИ ВИРІШЕННЯ ЛЕКСИЧНОЇ БАГАТОЗНАЧНОСТІ. ВИКОРИСТАННЯ WORDNET ДЛЯ ВИРІШЕННЯ ПРОБЛЕМ БАГАТОЗНАЧНОСТІ

Здійснено аналіз-ознайомлення за найпоширенішими методами та алгоритмами вирішення лексичної багатозначності. Розглянуто можливість їх практичної реалізації.

Ключові слова: вирішення лексичної багатозначності, методи навчання з вчителем та без вчителя, методи на основі знань, інформаційні ресурси, контекст, WordNet, лексична семантика.

This paper deals with the insight and analysis of the most widespread Word Sense Disambiguation methods and algorithms. The possibility of their practical implementation has been considered.

Keywords: Word Sense Disambiguation, supervised and unsupervised learning methods, knowledge-based methods, information resources, context, WordNet, lexical semantics.

1. Постановка проблеми

Полісемія – це явище багатозначності слова. Мовленнєва багатозначність безпосередньо пов’язана з поліфункціональністю контексту щодо певного слова. Для розуміння тексту чи навіть окремих висловлювань дуже важливим є визначення правильного значення слів відносно контексту. Вирішення лексичної багатозначності (Word Sense Disambiguation, WSD) – це завдання опрацювання природної мови, яке полягає в виборі значення (або сенсу) багатозначного слова чи словосполучення залежно від контексту.

Наукові дослідження з вирішення лексичної багатозначності перебувають у полі зору прикладної та комп'ютерної лінгвістики достатньо давно і мають багату історію, але повного вирішення проблема поки не отримала, оскільки на шляху успішного вирішення стоїть багато перешкод, безпосередньо пов'язаних з особливостями людської мови .

Отже, проблема вирішення лексичної багатозначності є насправді актуальною у сфері комп’ютерної лінгвістики, оскільки її розв’язання допоможе значно покращити ефективність опрацювання природної мови, що призведе до кращого вирішення завдань у цій сфері. Виділяють три основні підходи до вирішення проблем багатозначності: методи навчання з вчителем, методи навчання без вчителя, а також методи на основі знань [5]. Кожен підхід має свою специфіку та способи практичної реалізації, які будуть розглядатися у статті.

2. Цілі статті

Мета цього дослідження – аналіз методів WSD і постановка завдання їх практичної реалізації.

Отже, цілями статті є:

- ознайомлення з проблемою вирішення лексичної багатозначності;
- дослідження вищезгаданих методів вирішення завдань цього напряму;
- аналіз можливості використання лексичної бази WordNet для вирішення лексичної багатозначності;
- визначення необхідних заходів для практичної реалізації методів WSD.

3. Основний матеріал

3.1. Загальні відомості про дослідження в сфері WSD

3.1.1. Опис проблематики WSD

Вирішення лексичної багатозначності (WSD) передбачає пошук асоціації, яка викликана певним словом в тексті чи дискурсі із значенням, яке вирізняється від інших можливих значень цього слова. Тому шлях до вирішення завдання обов'язково складається з декількох кроків. Перший передбачає визначення всіх можливих потенційних значень кожного слова, що стосуються тексту або дискурсу. Другий крок включає засоби визначення правильного значення при кожній появі слова в контексті. Вся робота щодо вирішення неоднозначності передбачає врахування контексту, в якому вжите слово, і використання даних з зовнішніх джерел інформації. Також передбачений і третій крок: комп'ютер повинен навчитися співвідносити значення слова і саме слово в контексті, використовуючи машинне навчання або правила, які створив дослідник [6].

Людська мова неоднозначна, тому багато слів можуть інтерпретуватись по-різному. Для прикладу, розглянемо два речення:

(1) I can hear **bass** sounds. (2) They like grilled **bass**.

Слово *bass* в двох реченнях має різні значення: низькочастотні тони і вид риби.

У більшості випадків людина не думає про неоднозначності в мові, а ось комп'ютер повинна обробляти неструктуровану текстову інформацію і перетворювати її на структуровані дані, які аналізуються для визначення основного значення слова чи вислову. Саме визначення значення за допомогою комп'ютера і є основою WSD.

Успішність підходів до вирішення WSD залежить від ряду факторів. По-перше, завдання може бути сформульоване по-різному залежно від фундаментальних питань, наприклад, підхід до подання значення, рівень структурованості представлення значень слова, тексти на визначену тематику та тематично не обмежені, множина (кількість) слів для опрацювання і т.д.

По-друге, WSD працює на основі бази знань. Будь-який процес WSD містить множину слів (наприклад, речення або збірка слів) та техніку, яка використовує одне або декілька джерел знань, для визначення правильного значення слова відповідно до контексту. Джерела знань можуть різнитися, починаючи від корпусів текстів до більш структурованих джерел, таких як електронні словники, семантичні мережі тощо. Без баз знань було б неможливо як для людей, так і для машин визначити значення слова.

На жаль, створення баз знань власноруч вимагає багато часу та зусиль, а також структурних змін кожного разу після зміни сценарію WSD. Ці проблеми і формують сферу досліджень у WSD.

3.1.2. Завдання та основні елементи WSD

Зазвичай методи WSD мають однакову основу для опрацювання. Для того, щоб приступати до виконання завдання WSD (визначення значення усіх або декількох слів з тексту), потрібно спершу знехтувати знаками пунктуації. Тоді перед нами постане текст як послідовність слів. WSD можна розглядати як класифікатор: можливі значення слова – це класи, а автоматичний метод класифікації повинен фіксувати появу певного слова та відносити до одного чи декількох класів на основі контексту і зовнішніх баз знань.

Розрізняють два основних варіанти головних завдань WSD:

- Лексичний зразок, де система повинна опрацьовувати визначену множину слів, які можуть зустрітись лише один раз на речення.
- WSD для усіх слів, де система повинна опрацьовувати слова відкритих класів (дієслова, іменники, прикметники та прислівники).

Можемо виділити основні елементи WSD: вибір значень слова (класів), використання зовнішніх баз знань, представлення контексту, вибір методу класифікації.

Зовнішні бази знань

Сьогодні не існує необхідної бази знань, тому використовують різного виду інформаційні ресурси. Інформаційні ресурси надають інформацію, яка допомагає визначити значення слова. Розрізняють структуровані та неструктуровані ресурси.

До структурованих ресурсів відносять:

• *Тезаурус* надає інформацію про зв'язки між словами: синонімія, антонімія тощо. Найпоширенішим тезаурусом у сфері WSD є Roget's International Thesaurus (Roget, 1911), використовують також Macquarie Thesaurus (Bernard, 1986).

• *Електронні словники* стали популярним джерелом знань для обробки природної мови з 1980-х рр.. Серед них: Collins English Dictionary, Oxford Advanced Learner's Dictionary of Current English, Oxford Dictionary of English (Soanes and Stevenson, 2003), а також Longman Dictionary of Contemporary English (LDOCE) (Proctor 1978).

• *Онтології* спеціалізуються на концептуалізації спеціальних тематичних доменів, включаючи, зазвичай, таксономію та множину семантичних зв'язків. Наприклад, WordNet можна розглядати як онтологію.

До неструктурованих ресурсів відносять:

• *Корпус* – набір текстів, які використовуються у навчальних моделях. Корпус може бути розміченим або нерозміченим. Оскільки ми маємо справу з розрізненням значень слова, нас цікавить саме семантична розмітка. Тобто розміченими вважатимемо корпуси, які містять опрацьовані значення. Обидва види корпусів використовуються в WSD. Найбільш продуктивно їх використовують у підходах, що передбачають навчання з учителем та без учителя.

- Неопрацьовані корпуси: Brown Corpus (1961), the British National Corpus (100 млн. слів письмових та усних зразків англійської мови), Wall Street Journal (30 млн. слів) та ін.
- Розмічені корпуси: SemCor – найбільший і найпопулярніший розмічений корпус, містить 352 тексти і приблизно 234 тис. зафіксованих значень, MultiSemCor – анлійсько-італійський паралельний корпус з опрацьованими значеннями за допомогою WordNet англійської та італійської мов та багато інших.

• *Набори колокацій* фіксують найчастіші випадки поєднання одного слова з іншими. Приклади колокацій доступні в таких джерелах: The British National Corpus collocations, the Collins Cobuild Corpus Concordance та ін..

• *Інші ресурси*, такі як: списки частотних слів (наприклад, списки функціональних беззмістовних слів, таких як артиклі a, an, the), доменні мітки.

Представлення контексту

Текст – це неструктуроване джерело інформації, тому для зручності використання його, як правило, перетворюють на структурований формат. Для цього попередньо обробляють вхідний текст, що передбачає такі етапи:

• токенізація – поділ тексту на токени (зазвичай слова);

• морфологічний аналіз – визначення частин мови (наприклад, "the/DT bar/NN was/VBD crowded/JJ", де DT, NN, VBD, JJ позначають артикль, іменник, дієслово та прикметник);

• лематизація – визначення початкової морфологічної основи слова (наприклад, was → be, bars → bar);

• чанкінг – поділ тексту на синтаксичні частини (наприклад, поділ [the bar was crowded] на $[\text{the bar}]_{NP}$ $[\text{was crowded}]_{VP}$, відповідно іменникову та дієслівну фрази);

• синтаксичний аналіз речень – побудова синтаксичного дерева відповідно до структури речення.

Послідовність та результати попередньої обробки тексту наведено на рис. 1.

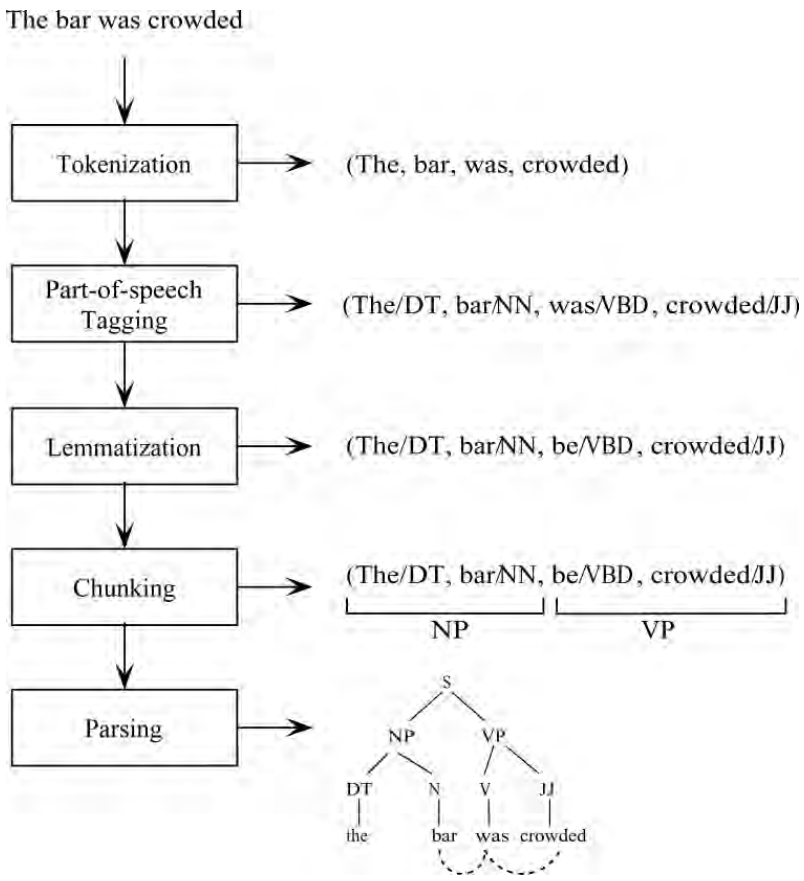

Рис. 1. Приклад обробки тексту (речення)[5]

Вибір методу

Як вже раніше згадувалось, існують різні підходи до вирішення проблем багатозначності. Досліджуючи різноманітні методи вирішення проблеми багатозначності, виділяють три основні підходи: методи навчання з вчителем та без вчителя, а також методи на основі знань. Вибір методу залежить від сформульованого завдання або мети проведення дослідження [5].

3.2. Методи вирішення багатозначності

3.2.1. Навчання із вчителем

Такі методи навчання використовують розмічені корпуси. Результатом цього методу будуть сформовані правила.

Для прикладу розглянемо процедуру опрацювання одного слова w з m значеннями $w_1,\ldots, w_m$. Визначаємо усі контексти корпусу, в яких зустрічається слово: $c_1,\ldots,c_m$. Також знадобиться словник корпусу: слова $v_1,\ldots,v_p$, які зустрічаються в контексті з w.

Ідея включає вибір значення, яке максимально підходить до умови: w' найкраще значення в контексті c, якщо, для будь-якого $w_к$ (крім w'), $P(w' \mid c) \geq P(w_k \mid c)$.

У таких методах часто використовується теорема Байєса:

$$P(w_k \mid c) = P(c \mid w_k) * P(w_k) / P(c) \quad (1)$$

Наївний класифікатор Байєса інформує, що елементи контексту незалежні. Якщо $c = v_a\, v_b \ldots v_z$, можна приблизно визначити $P(w_k \mid c)$:

$$P(v_a \mid w_k) * P(v_b \mid w_k) * \ldots * P(v_z \mid w_k) \quad (2)$$

Можливість того, що елемент контексту v_i, значення w_k, визначені підрахунком одночасної появи і v_i та w_k у корпусі. Те ж саме з w_k.

$$P(v_i \mid w_k) = C(v_i, w_k) / C(w_k) \quad (3)$$

$$P(w_k) = C(w_k) / C(w) \quad (4)$$

$P(c)$ залишається незмінним, тому його можна проігнорувати, якщо ми просто хочемо збільшити значення $P(w_k \mid c)$. Можна зазначити, що із збільшенням логарифма значення має той самий ефект, що й збільшення самого значення. Щоб знайти w_k, що збільшується $P(v_a \mid w_k) * P(v_b \mid w_k) * ... * P(v_z \mid w_k) * P(w_k)$ те є саме, що:

$$log\ P(v_a \mid w_k) + log\ P(v_b \mid w_k) ... + log\ P(v_z \mid w_k) + log\ P(w_k) \quad (5)$$

Алгоритм для методу навчання із вчителем містить такі кроки.

Крок 1 (тренування)

- Для усього контексту пари слово-значення v_i - w_k, необхідно підрахувати (3).
- Для усіх значень w_k, необхідно підрахувати (4).

Крок 2 (тестування)

- Дано слово w, вибір w_k, яке збільшуємо (5).

Розглянемо приклади елементів контексту, які допомагають визначити значення слова:

- Слово "bank (банк)" у значенні фінансової установи використовується у контексті з словами interest (відсоток), teller (касир), account (рахунок), ...
- Слово "bank (берег)" у значенні берега річки визначається за словами water (вода), river (ріка), right (правий), ... [9]

Одним із алгоритмів методів навчання з вчителем є побудова списку рішень. Список рішень – це впорядкований набір правил та дерева рішень. Правила можуть мати вигляд “if-then-else (якщо-то-або)”. Впорядкування правил залежить від підрахунку «балів». Значення, яке набуває найбільшу кількість балів, визначаємо так:

$$S = \arg\max_{S_i \in SensesD(w)} score(S_i) \quad (6)$$

$$score(S_i) = \max_f \log\left(\frac{P(S_i \mid f)}{\sum_{j \neq i} P(S_i \mid f)}\right),$$

де S_i – значення, f – властивості значення, $P(S_i \mid f)$ – ймовірність появи значення S_i та f.

Наприклад, для слова bank може бути побудований такий список рішень (рис. 2):

Ознака	Можливе значення	Оціночні бали
account with bank	Bank/Finance	4.83
stand/V *on*/P … bank	Bank/Finance	3.35
bank *of blood*	Bank/Supply	2.48
work/V … bank	Bank/Finance	2.33
the left/J bank	Bank/River	1.12
of the bank	-	0.01

Рис. 2.Список рішень

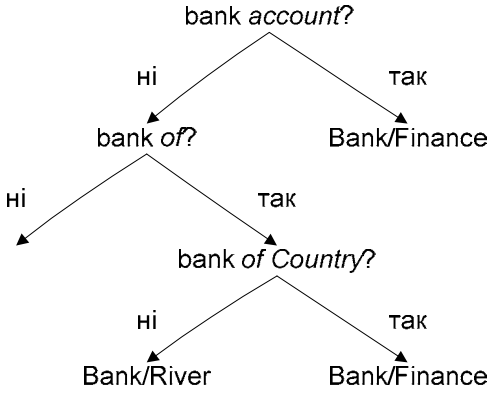

Рис. 3. Дерево рішень

Дерево рішень (рис. 3) – це модель представлення правил класифікації у вигляді дерева, яка рекурсивно розподіляє множину даних для тренування системи. Кожен внутрішній вузол дерева рішень представляє перевірку властивості значення, а кожна гілка надає результат перевірки. Визначення можливого значення завершується із досягненням кінцевого вузла [5].

3.2.2. Навчання без вчителя

Ідея такого підходу полягає у тому, що одне й те саме значення слова матиме однакові сусідні слова – елементи контексту. Так можна визначити значення із вхідного тексту за допомогою кластеризації слів і класифікації нових слів до сформованих кластерів. Методи не залучають розміченого тексту, електронних словників, тезаурусів тощо. Хоча недоліком таких методів є відсутність таких даних про значення слів, які, наприклад, є у словниках. Основними методами такого підходу є: кластеризація контексту та слів та графи співвідношень.

Кластеризація контексту

Кожна поява потрібного слова в корпусі представлена контекстним вектором. Вектор включає усі значення слова. Контекстні вектори групуються у кластери, кожен з яких визначає значення слова. Далі відбувається пошук найбільш схожих між собою кластерів.

Наприклад, для слова restaurant (ресторан) = (210, 80) та money (гроші) = (100, 250); вказані у дужках значення – це кількості появ слова зі словом food (їжа) перше значення , а друге – зі словом bank (банк).

Співвідношення між двома (v та w) словами визначається геометрично:

$$sim(\upsilon, w) = \frac{\upsilon * w}{|\upsilon||w|} = \frac{\sum_{i=1}^{m} \upsilon_i w_i}{\sqrt{\sum_{i=1}^{m} \upsilon_i^2 \sum_{i=1}^{m} w_i^2}}, \quad (7)$$

де m – кількість властивостей кожного вектора.

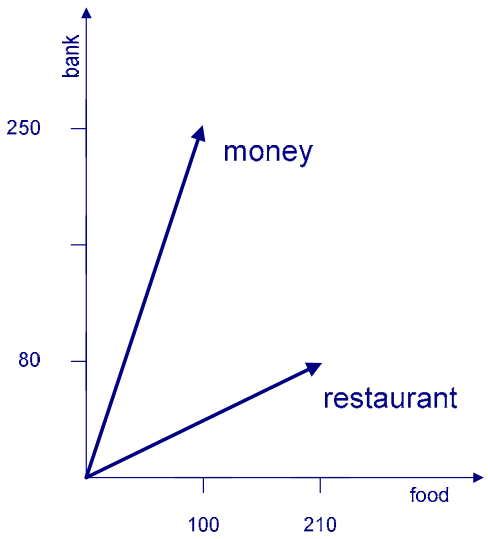

Рис. 4. Приклад векторів

Кластеризація слів

Кластеризація слів складається із визначення слів $W = (w_1, \ldots, w_k)$, схожих на основне слово w_0. Подібність слів w_1 та w_0 визначається на основі інформації про їх властивості, які відображаються у синтаксичних залежностях корпусу (наприклад, суб'єкт-дієслово, дієслово-об'єкт,

прикметник-іменник і т.д.). Чим більше залежностей між двома словами, тим більше інформації. W – список подібних слів, впорядкованих рівнем схожості до w_0. Дерево схожостей T складається з одного вузла w_0. Для кожного $i \in \{1, \ldots, k\}$, $w_i \in W$ додається елемент w_j – найближче слово до w_i серед $\{ w_0, \ldots, w_{i-1} \}$.

Визначення кластерів допомагає визначити найменш можливі значення слова.

Графи співвідношень

Цей метод передбачає побудову графів вигляду $G = (V, E)$, де V представляє слова в тексті, а дуги E зв'язують пари слів, зважаючи на синтаксичні відношення. Основним алгоритмом такого методу вважають HyperLex [10]. Приклад результату такого алгоритму представлено на рис. 5.

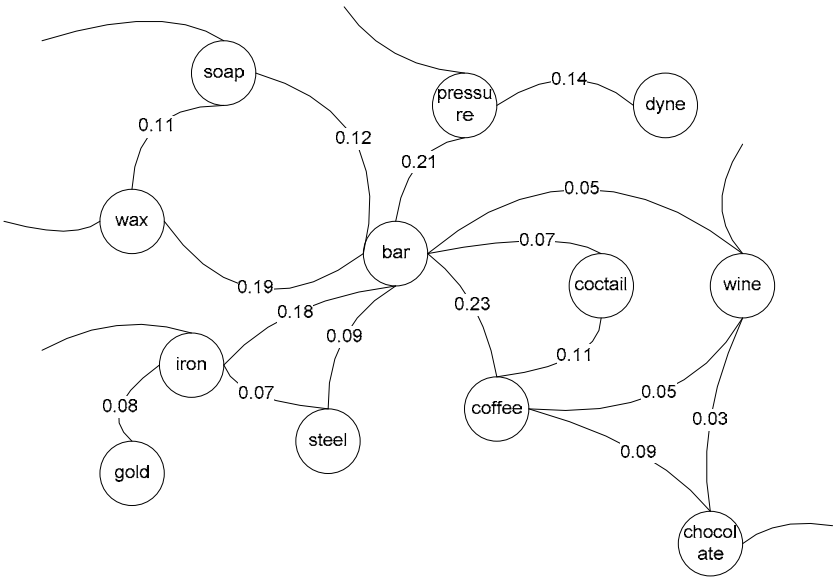

Рис. 5. Частина графу для слова bar

3.2.3. Методи на основі знань

Метою таких методів є використання бази знань. Враховуючи відсутність такої бази, використовують інформаційні ресурси: словники, тезауруси, онтології, колокації тощо. Розглянемо декілька з них.

Методи з використанням словника

Для кожного можливого значення слова w_k корпусу повинно бути визначення із словника. Визначення представлене у вигляді невпорядкованого набору слів (можливі повторення). D_{Vij} – визначення до значення j кожного словникового слова v_i, E_{Vi} – сума усіх D_{Vij}.

Алгоритм визначення значення представлений у [11]:

- Вибираємо слово w і його контекст $c = v_a\, v_b \ldots v_z$.
- Підраховуємо суму $E_{Va} \cup E_{Vb} \cup \ldots \cup E_{Vz} = D_C$.
- Підраховуємо бали при кожному збігу між D_{Wk} та D_C.
- Обираємо найбільш можливе w_k.

Методи з використанням тезаурусу

Припустимо, що кожне значення слова w_k має семантично розрізнюючий тег $t(w_k)$ (наприклад, номер синсету у WordNet або посилання у Roget's Thesaurus). Визначається відстань між словом v і семантичним тегом t:

$d(t, v) = 1$, якщо це тег слова v та $d(t, v) = 0$, якщо це тег не слова v

Простий алгоритм визначення значення з використанням семантичних тегів:

- Вибираємо слово w і його контекст $c = v_a\ v_b\ ...\ v_z$.
- Підраховуємо бали кожного значення w_k оцінкою $d(t(w_k), v_a) + d(t(w_k), v_b) + ... + d(t(w_k), v_z)$.
- Обираємо w_k з найбільшою кількістю балів.

Існує також багато інших алгоритмів визначення наймовірнішого значення слова. Найпопулярнішими ресурсами, які використовуються під час вирішення таких завдань, є WordNet та SemCor.

SemCor – це частина браунівського корпусу, де слова морфологічно опрацьовані, а значення слів взяті з системи WordNet. SemCor складається з 352 текстів: 186 текстів з використанням слів відкритих класів (іменники, дієслова, прикметники та прислівники), 166 текстів містять семантичні характеристики лише дієслів.

WordNet – це електронний лексикон англійської мови у вигляді множин синонімів – синсетів. Цей ресурс містить понад 155 тис. слів, організованих у близько 177 тис. синсетів.

З усіх трьох підходів найефективнішими ми вважаємо методи на основі знань, оскільки інформаційні ресурси забезпечують величезний спектр інформації, яку можна використовувати для вирішення лексичної багатозначності. Найбільше можливостей для роботи у цій сфері надає WordNet.

У наступному розділі ми детальніше розглянемо можливості використання WordNet для розв'язання проблеми розрізнення значень слів.

3.3. Застосування WordNet у вирішенні проблем багатозначності

3.3.1. Загальні відомості про WordNet

WordNet – це мережа слів, спроектована відповідно до сучасних психолінгвістичних теорій про людську лексичну пам'ять. Англійські іменники, дієслова, прикметники і прислівники утворюють синонімічні набори (синсети), кожен з яких представляє один лексичний концепт. Ці набори пов'язані концептуально-семантичними та лексичними зв'язками.

WordNet розробили у Когнітивній науковій лабораторії при Прінстонському університеті під керівництвом професора Джорджа А. Міллера (http://www.cogsci.princeton.edu/). WordNet вважається найважливішим ресурсом для досліджень в комп'ютерній лінгвістиці, аналізі тексту та в багатьох суміжних галузях. Його розроблено відповідно до сучасних фізіологічних та обчислювальних теорій лексичної пам'яті людини. WordNet зовні нагадує тезаурус, тому що він групує слова залежно від їх значення. Однак є деякі важливі відмінності. По-перше, у WordNet містяться не лише зв'язки між словоформами, а й конкретні значення слів. У результаті слова, що перебувають в безпосередній близькості одне від одного в мережі, семантично розрізнені. По-друге, WordNet, на відміну від тезауруса, позначає семантичні зв'язки між словами [7].

Кожне слово в структурі WordNet морфологічно позначено. Синсет можна розглядати як множину значень слів, які виражають одне поняття. Наприклад, поняття «cat» (кіт)» пов'язане з десятьма різними синсетами, і його можна представити як:

$$Senses_{WN}(car_{n,v}) = \left\{ \begin{array}{c} \{\mathbf{cat}^1_{\mathbf{n}},\ \text{true cat}^1_{\text{n}}\}, \\ \{\text{guy}^1_{\text{n}},\ \mathbf{cat}^2_{\mathbf{n}},\ \text{hombre}^1_{\text{n}},\ \text{bozo}^2_{\text{n}},\ \text{sod}^4_{\text{n}}\}, \\ \{\mathbf{cat}^3_{\mathbf{n}}\}, \\ \{\text{kat}^1_{\text{n}},\ \text{khat}^1_{\text{n}},\ \text{qat}^1_{\text{n}},\ \text{quat}^1_{\text{n}},\ \mathbf{cat}^4_{\mathbf{n}},\ \text{Arabian tea}^1_{\text{n}},\ \text{African tea}^1_{\text{n}}\}, \\ \{\text{cat} - \text{o}' - \text{nine} - \text{tails}^1_{\text{n}},\ \mathbf{cat}^5_{\mathbf{n}}\}, \\ \{\text{Caterpillar}^2_{\text{n}},\ \mathbf{cat}^6_{\mathbf{n}}\}, \\ \{\text{big cat}^1_{\text{n}},\ \mathbf{cat}^7_{\mathbf{n}}\}, \\ \{\text{computerized tomography}^1_{\text{n}},\ \text{computed tomography}^1_{\text{n}},\ \text{CT}^2_{\text{n}},\ \text{computerized axial tomography}^1_{\text{n}}, \\ \text{computed axial tomography}^1_{\text{n}},\ \mathbf{CAT}^8_{\mathbf{n}}\}, \\ \{\mathbf{cat}^1_{\mathbf{v}}\}, \\ \{\text{vomit}^1_{\text{v}},\ \text{vomit up}^1_{\text{v}},\ \text{purge}^6_{\text{v}},\ \text{cast}^{11}_{\text{v}},\ \text{sick}^1_{\text{v}},\ \mathbf{cat}^2_{\mathbf{v}},\ \text{be sick}^1_{\text{v}},\ \text{disgorge}^2_{\text{v}},\ \text{regorge}^1_{\text{v}},\ \text{retch}^1_{\text{v}}, \\ \text{puke}^1_{\text{v}},\ \text{barf}^1_{\text{v}},\ \text{spew}^3_{\text{v}},\ \text{spue}^2_{\text{v}},\ \text{chuck}^4_{\text{v}},\ \text{upchuck}^1_{\text{v}},\ \text{honk}^4_{\text{v}},\ \text{regurgitate}^4_{\text{v}},\ \text{throw up}^1_{\text{v}}\} \end{array} \right\}$$

Кожне слово однозначно ідентифікує один синсет. Наприклад, cat_n^2 відповідає синсет $\{\text{guy}_\text{n}^1, \text{cat}_\text{n}^2, \text{hombre}_\text{n}^1, \text{bozo}_\text{n}^2, \text{sod}_\text{n}^4\}$. Для кожного синсету WordNet представляє таку інформацію:

- *Глоса* – текстове визначення синсету з множиною прикладів використання слова у реченні.
- *Лексичні та семантичні зв'язки*, які пов'язують значення та синсети пари слів. Семантичні зв'язки створюють в синсетах цілісність, лексичні зв'язки зв'язують значення слова в синсетах.

Серед лексичних зв'язків виділяють:

- *Антонімія*: Х антонім У, якщо він виражає протилежне поняття.
- *Слова-відношення*: слова, що стосуються одного поняття, навіть якщо вони мають різні частини мови (наприклад, прикметник dental співвідноситься з іменником tooth).
- *Номіналізація*: утворення іменника з дієслова (наприклад, serve та service).

Серед семантичних зв'язків розрізняють:

- *Гіперніми*: У гіпернім Х, якщо кожне Х – це вид У (наприклад, motor vehicle – гіпернім слова car). Гіперніми та гіпоніми є для іменникових та дієслівних синсетів.
- *Гіпоніми:* протилежне до гіпернімів поняття (наприклад, car – гіпонім motor vehicle).
- *Мероніми*: У меронім до Х, якщо У частина Х (наприклад, flesh меронім fruit). Властиве лише іменниковим синсетам.
- *Голоніми*: протилежне до меронімів поняття (наприклад, fruit голонім flesh).
- *Імплікація*: дієслово У імплікація до Х, якщо, виконуючи Х, повинен відбуватися У (наприклад, snore передбачає sleep).
- *Тотожність*: прикметник Х і прикметник У ідентичні (наприклад, beautiful та pretty).
- *Ознака*: іменник Х є ознакою, для якої прикметник У виражає значення (наприклад, hot ознака temperature).

Як бачимо, WordNet надає дуже різноманітну інформацію про слова. Усі вище перелічені типи зв'язків можна використати для вирішення лексичної багатозначності. Так, зважаючи на високу продуктивність та популярність серед дослідників, WordNet можна вважати стандартом англійської мови у вирішенні проблем сфери WSD [5].

Далі конкретніше розглянемо можливості практичного використання WordNet для WSD.

3.3.2. Практичне застосування WordNet

Існує багато методів та алгоритмів використання WordNet самостійно чи з використанням інших ресурсів, наприклад, корпусів текстів. Розглянемо декілька алгоритмів з використанням WordNet. Як раніше згадувалось, провідну роль у визначенні значення слова відіграє контекст.

Наприклад, дано речення «Tax revision **bills** were passed». У цьому випадку цільове слово – *bills* і чотири елементи контексту – tax, revision, were, passed. Основна ідея алгоритму полягає у підрахунку балів, які «підтверджуватимуть» одне зі значень, зважаючи на контекст:

$$, score_{s_i} = \sum_{j,k} \max_k relatedness(s_i, s_{jk}) , \quad (8)$$

де s_{jk} – *k*-те значення *j*-го контекстного слова.

Обирається значення з найбільшою кількістю балів [9].

Інший підхід до використання WordNet також передбачає врахування контексту, як єдиного показника значення слова. Алгоритм передбачає врахування і морфологічної розмітки. Розглядаються три типи груп: група b1 складається з елементів контексту; домени для кожного слова з b1 сортуються за частинами мови до групи b2. Домени відповідно до частини мови цільового слова входять до групи b3. Домени групи b3 порівнюються з доменами контекстних слів. Домен основного слова, який найбільше збігається з доменами інших слів, стає доменом тексту. Значення, яке несе цей домен, стає правильним для основного слова.

Припустимо, що $(w_1, w_2, w_3,..,w_n)$ – слова контексту з визначеними частинами мови, які відносяться до групи b1. Група b2 складається з $(d_1, d_2, d_3,\ldots,d_n)$ – набору доменів відповідно до слів контексту і частин мови. Кожний набір d_i містить усі можливі домени. Група b3 складається з доменів основного слова.

Для прикладу розглянемо речення:

«The **virus** infected all files on the hard disk».

Після здійснення морфологічної розмітки воно матиме такий вигляд:

The/DT virus/NN infected/VBD all/DT files/NNS on/IN hard_disk/NN ./.

Для обробки обираються основні слова контексту, тобто virus, infected, files та disk. Основним словом обираємо *virus*. Групи b1, b2, b3 матимуть такий вигляд:

b1	b2				b3
	Virus *(noun sense)* Target word	Infected *(verb sense)*	Files *(noun sense)*	Hard_disk *(noun sense)*	
Virus *(noun sense)*	{01331343} <animal>	{00089502} <body>	{06520807} <communication>	{03497643} <artifact>	{01331343} <animal>
Infected *(verb sense)*	{14031349} <state>	{00088465} <body>	{08445713} <group>		{14031349} <state>
Files *(noun sense)*	{06597992} <communication>	{02586322} <social>	{03342085} <artifact>		{06597992} <communication>
Disk *(noun sense)*		{00606893} <cognition>	{03341784} <artifact>		

Рис. 7. Вміст основних груп b1, b2, b3

Кожний домен з b3 порівнюється з кожним доменом кожного з контекстних слів. Домен з b3, який максимально збігається з доменами контекстних слів, і стає доменом тексту. Значення обраного домену відповідає значенню основного слова. Значення домену <communication> з групи b3 стає значенням основного слова.

Наступний підхід до вирішення багатозначності передбачає використання принципів ієрархії у системі WordNet. Залучення лексичних та семантичних відношеннь удосконалюють вирішення багатозначності.

А. Використання гіпернімів.

Наприклад, речення «He ate many **dates**.» Після визначення частин мови: «He/ PRP ate/VBD many/ DT dates/ NNS». Слово date має 8 значень – гіпернімів:

1. (503) date, day of the month
2. (119) date -- (a particular day specified as the time something happens)
3. (104) date, appointment, engagement
4. (55) date, particular date
5. (37) date -- (the present; "they are up to date")
6. (29) date, escort -- (a participant in a date)
7. (26) date -- (the particular day, month, or year)
8. (20) date -- (sweet edible fruit)

Після додаткових операцій алгоритм визначить, що date (фінік) – вид їстівного фрукту.

Б. Використання меронімів та голонімів.

Наприклад, речення «The **trunk** is the main structural member of a tree that supports the branches». Після визначення частин мови: «The/DT trunk/NN is/VBZ the/DT main/JJ structural_member/NN of/IN a/Z tree/NN that/WDT supports/ VBZ the/DT branches/NNS». Слово *trunk* має значення:

1. {13186713} trunk#1 PART OF: {13124818} <noun.plant>[20] S: (n) tree#1,
3. {05557463} trunk#3 PART OF: {05223633} <noun.body>[08] S: (n) body#1,
4. {03701391} trunk#4 PART OF: {02961779} <noun.artifact>[06] S: (n) car#1,
5. {02455598} trunk#5 PART OF: {02506148} <noun.animal>[05] S: (n) elephant#1
6. {02455598} trunk#5 PART OF: {02507401} <noun.animal>[05] S: (n) mammoth#1

Контекст містить слово *tree*, алгоритм визначить правильне значення - sense #1.

Використовуючи можливості WordNet, використовуючи усі можливі зв'язки між синсетами, можна досягнути хорошого результату у визначенні правильного значення слова у тексті [3].

Висновки

Сфера вирішення лексичної багатозначності має багату та продуктивну історію, але до цього часу залишається актуальною та активно розвивається. Періодичні конференції та незалежні дослідники працюють над вдосконаленням шляхів вирішення основних завдань цієї галузі, адже WSD має надзвичайний вплив на такі завдання опрацювання природної мови, як інформаційний пошук, машинний переклад, видобування інформації, контент-аналіз, лексикографію, семантику веб-документів і т.д.

У цій статті ми дослідили проблему вирішення лексичної багатозначності та проаналізували основні методи її розв'язання, кожен з яких пропонує своє рішення проблеми та шляхи вдосконалення розв'язків, і, звичайно, кожен з них має як переваги, так і недоліки, які потребують подальшого доопрацювання.

Серед трьох основних підходів до вирішення лексичної багатозначності ми обрали методи на основі знань. Ми детально дослідили сучасні інформаційні ресурси і дійшли висновку, що мережа WordNet надає найбільше можливостей для автоматичного розрізнення значень слів. Алгоритм підрахунку балів та використання принципів ієрархії WordNet дають можливість створити систему автоматичного вирішення лексичної багатозначності.

1. Разрешение лексической многозначности. – Available from: http://ru.wikipedia.org/wiki/ Разрешение_лексической_многозначности. 2. Kikas T. Word Sense Disambiguation WordNet::SenseRelate:: AllWords / T. Kikas, M. Treumuth. – 2007. – Available from: http://math.ut.ee/~treumuth/NLP/semantics2.pdf 3. Kolte S. WordNet: A Knowledge Source for Word Sense Disambiguation / S. Kolte, S. Bhirud. – India. – 2009. – Available from: http://www.academypublisher.com/ijrte/vol02/no04/ijrte0204213217.pdf. 4. Mocian H. Survey of Word Sense Disambiguation / H. Mocian. - Available from: http://www.horatiumocian.com/papers/ Word_Sense_Disamb_Survey.pdf. 5. Navigli R. Word Sense Disambiguation: A Survey / R. Navigli. – 2009. – Available from: http://www.dsi.uniroma1.it/~navigli/pubs/ACM_Survey_2009_Navigli.pdf 6. Palta E. Word Sense Disambiguation / E. Palta. – 2006-2007. – Available from: http://citeseerx.ist.psu.edu/viewdoc/ download?doi=10.1.1.102.2419&rep=rep1&type=pdf. 7. What is WordNet? – Available from: http://wordnet.princeton.edu/. 8. WordNet-based Algorithm for Word Sense Disambiguation. – Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi= 10.1.1.47.6728&rep=rep1&type=pdf. 9. Word Sense Disambiguation. – Available from: http://www.site.uottawa.ca/~szpak/teaching/5386/handouts/WSD-memo.pdf. 10. J. V´eronis. 2004. Hyperlex: lexical cartography for information retrieval. Computer Speech & Language,. 18(3):223-252. 11. Michael Lesk, Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone, ACM Special Interest Group for Design of Communication Proceedings of the 5th annual international conference on Systems documentation, p. 24 – 26, 1986.