лексикографічного моделювання) // Наукові записки. – Вип. 126. – Серія: Філологічні науки (мовознавство). – Кіровоград: РВВ КДПУ ім. В. Винниченка, 2014. – С. 242–248.
16. Сніжко Н. В. Зведений словник лексики української мови кінця XVIII – поч. XXI ст. в інтегральних дослідженнях та компаративістиці // Компаративні дослідження слов'янських мов і літератур. – К.: Київський нац. ун-т ім. Т. Шевченка, 2015. – Вип. 27. – С. 174–185.
17. Сніжко Н. В. Ідеографічна група ЛЮДИНА в українській мові другої половини XX ст. (структурно-функціональний і аксіологічний аспекти) // Компаративні дослідження слов'янських мов і літератур. – К.: Київський нац. ун-т ім. Т. Шевченка, 2016. – Вип. 29. – С. 120–129.
18. Сніжко Н. В. «Ідеографічний тезаурус» як інформаційно-довідкова система при вивченні закономірностей структурно-функціональної організації лексики / Н. В. Сніжко, М. Д. Сніжко // Мовознавство. – 1996. – № 4–5. – С. 23–28.
19. Сніжко Н.В. Категоризація знань про світ і мову в інтегральних лексикографічних системах // Вісник Прикарпатського національного університету. Філологія. – Випуск 36–37. – Івано-Франківськ: Видавництво Прикарпатського національного університету, 2012. – С. 28–32.
20. Сніжко Н. В. Проблеми сучасної інтегральної лексикографії // Слово и словарь = Vocabulum et vocabularium : сб. науч. тр. / ГрГУ им. Я. Купалы ; редкол.: Л. В. Рычкова (гл. ред.), Т. Ройтер, В. В. Дубичинский [и др.]. – Гродно : ГрГУ, 2013. – 203 с.
21. Сніжко Н. В. Системний підхід до вивчення динаміки лексичного складу української мови кінця XVIII – поч. XXI ст. // Українська мова. – 2013. – № 3. – С. 110–127.
22. Сніжко Н. В. Структурна, функціональна та хронологічна параметризація лексики у зведеному словнику, тезаурусі й електронній картотеці // Українська мова. – 2016. – № 4. – С. 87–102.
23. Сніжко Н. В. Українська ідеографія: історія, сучасний стан та перспективи // Українська мова. – 2016. – № 3. – С. 28–43.
24. Струганець Л. В. Динаміка лексичних норм української літературної мови XX ст. – Тернопіль: Астон, 2002. – 352 с.
25. Тараненко О.О. Актуалізовані моделі в системі словотворення сучасної української мови (кінець XX – поч. XXI ст.). – К.: Видавничий дім Дмитра Бураго, 2015. – 248 с.
26. Тараненко О. О. Новий словник української мови (концепція і принципи укладання словника). – К.; Кам'янець-Подільський, 1996. – 172 с.
27. Українська і слов'янська тлумачна та перекладна лексикографія. Леонідові Сидоровичу Паламарчукові / Інститут української мови НАН України; відп. ред. І. С. Гнатюк. – К.: КММ, 2012. – 488 с.
28. Український орфографічний словник / уклали: В. В. Чумак [та ін.]; за ред. В. Г. Скляренка. – К.: Довіра, 2009, 1011 с. – (Словники України).
29. Украинский семантический словарь. Проспект / М.М. Пещак, Н. Ф. Клименко, Е. А. Карпиловская и др.; Отв. ред. М. М. Пещак. – К.: Наук. думка, 1990. – 264 с.

Розробка словника WordNet-Affect для української мови

Ольга Каніщева

к. т. н., доцент, доцент кафедри інтелектуальних комп'ютерних систем, Національний технічний університет «Харківський політехнічний інститут», Україна, E-mail: kanichshevaolga@gmail.com

Катерина Клименкова

студентка, Національний технічний університет «Харківський політехнічний інститут», Україна, E-mail: klim789@bk.ru

Катерина Юр'єва

студентка, Національний технічний університет «Харківський політехнічний інститут», Україна, E-mail: katy.yurieva@gmail.com

This article shows the creation process of WordNet-Affect-based lexical resource for the Ukrainian language. Authors translated WordNet-Affect into Ukrainian based on WordNet-Affect for Russian and Romanian languages. This resource can be used for automatic evaluation of news events, products, personalities, organizations, countries, etc. This semantic thesaurus can be used to detect and interpret thoughts, clustering texts of polar (positive or negative) opinions; segmentation of texts; forecasting opinions, based on the analyzed text.

Ключові слова — автоматична обробка природної мови, комп'ютерна лексикографія, Sentiment Analysis, WordNet-Affect, тональні словники.

I. Вступ

Із кожним днем кількість соціальних мереж, блогів і чатів зростає в Інтернет павутині, відповідно відбувається збільшення інформації, яка несе емоційне навантаження.

Завдання аналізу емоційного забарвлення текстів, розвиток методів фільтрації в мережі Інтернет набувають все більшої актуальності у зв'язку з величезною аудиторією мережі, зростаючим середнім часом перебування в ній та ін. Аналітика та моніторинг соціальних мереж становить величезний інтерес для соціологів, лінгвістів, психологів, маркетологів і державних структур. Саме ці факти роблять аналіз суб'єктивних текстів особливо актуальним [1].

Для вирішення завдань аналізу тексту, який емоційно забарвлений, в галузі обробки природної мови використовують методи контент-аналізу, загальна назва для яких – Sentiment Analysis (аналіз тональності тексту) [2].

Аналіз тональності зазвичай визначають як одну із задач класифікації, тобто мається на увазі, що ми можемо знайти і класифікувати тексти відповідно їх тональності, при цьому використовують такі інструменти обробки природної мови, як морфологічний та синтаксичний аналізатори, та ін. Існуючи підходи до Sentiment Analysis можна поділити на такі групи [3, 4]:

• підходи, які засновані на правилах;

• підходи, які використовують спеціалізовані словники;

• методи машинного навчання.

Підходи, засновані на словниках, використовують так звані тональні словники (affective lexicons) для аналізу тексту. У простому вигляді тональний словник представляє собою список слів зі значенням тональності для кожного слова.

Більшість лексичних ресурсів та програмних засобів для сентимент-аналізу було створено для англійської мови. Наприклад, WordNet-Affect, SentiWordNet, SentiNet та інші.

Але для східноєвропейських мов ситуація більш складна, вільно розповсюджуваних словників дуже мало. Тому, метою даної роботи є розробка спеціалізованого тонального словника для української мови, а саме семантичного тезаурусу WordNet-Affect, що дозволить використовувати цей ресурс для визначення тональності тексту.

II.Огляд існуючих тональних словників для української мови

Україномовних словників емоційної лексики, на жаль, дуже мало. Одним з них є "Український тональний словник" (https://github.com/lang-uk/tone-dict-uk).

Він містить 3,442 слів української мови, які мають не нейтральну тональність (-2, -1, 1, 2). Ці дані отримані з двох джерел: експертів та згенеровані автоматично за допомогою алгоритмів машинного навчання та також з використанням векторів слів word2vec та lex2vec. Після цього слова були оброблені людиною вручну.

У словнику всі слова нормалізовані та прислівники замінені на спільнокореневі прикметники.

Ще одним словником є Emotion Lexicon CPH (http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm). Він містить 8,903 емоційних слів і їх асоціації з вісьмома основними емоціями і почуттями (негативні та позитивні).

III.Словник WordNet-Affect

WordNet-Affect (wndomains.fbk.eu/wnaffect.html) – це семантичний тезаурус, в якому поняття, пов'язані з емоціями, («емоційні концепти», «Affective concepts») представлені за допомогою слів, що володіють емоційною складовою («емоційні слова», «Affective words») [5]. WordNet-Affect складається з такої підмножини сінсетов WordNet, де кожен сінсет, який відповідає певному емоційному концепту, може бути представлений за допомогою емоційних слів.

Таким чином, WordNet-Affect був створений на основі WordNet для англійської мови (також існують версії WordNet-Affect і для інших мов) шляхом вибору і віднесення наборів синонімів (сінсетов) до різних емоційних понять. Зокрема, сінсети дієслів, іменників, прикметників, прислівників, які представляють собою опис емоцій, були вручну розмічені за допомогою спеціальних емоційних міток (affective labels, A-labels). Ці емоційні мітки характеризують різні стани, що виражають настрої, емоційні відгуки, або ситуації, які викликають емоції. Приклади таких емоційних міток наведені в Табл. 1 [5].

ТАБЛ. 1

Приклади емоційних міток

ЕМОЦІЙНА МІТКА	ПРИКЛАД
Эмоція(emotion)	ім. гнів#1, дієсл. боятися#1 (fear)
Настрій (mood)	ім. ворожість#1 (animosisy), прикм. люб'язність#1J (amiable)
Особливість (trait)	ім. агресивність#1 (aggressiveness), прикм. який_змагається#1 (competitive)
Когнітивний стан (cognitivestate)	ім. замішання#2 (confusion), прикм. вражений#2 (dazed)
Фізичний стан (physicalstate)	ім.хвороба#1 (llness), прикм. знесилений#1 (allin)
Гедонічний сигнал (gedonicsignal)	ім. біль#3(hurt), ім. страждання#4 (suffering)
Ситуації, що викликають емоції (emotion-elicitingsituation)	ім. незручність#3 (awkwardness), ім. безпека#1 (outofdanger)
Емоційні відгуки (emotionalresponses)	ім. холодний піт#1 (coldsweat), дієсл. тремтіти#2 (tremble)
Вчинки (behaviour)	ім. злочин#1 (offense), прикм. загальмований#1 (inhibited)
Ставлення, позиція (attitude)	ім. нетерплячість#1 (intolerance), ім. оборонна позиція#1 (defensive)
Відчуття (sensation)	ім. холод#1 (coldness), дієсл. відчувати#3 (feel)

Також у WordNet-Affect використовуються додаткові емоційні мітки для того, щоб розділяти сінсети відповідно до їх емоційної валентності. Для цього визначаються чотири додаткові емоційні мітки:

позитивна, негативна, неоднозначна і нейтральна. Перша відповідає позитивним емоціям, які визначаються як емоційні стани, що характеризуються наявністю позитивних гедоністичних сигналів (або задоволення). Вона включає в себе такі сінсети, як радість#1 або захоплення#1. Аналогічно негативна мітка ідентифікує негативні емоції, які характеризуються негативними гедоністичними сигналами (або болю), наприклад, гнів#1 або печаль#1. Сінсети, що представляють емоційні стани, валентність яких залежить від семантичного контексту (наприклад, здивування#1) позначаються як неоднозначні. Сінсети, що визначають психологічні стани, і які завжди розглядаються як неоднозначні, але при цьому не характеризуються валентністю, є нейтральними.

Сінсети, помічені емоційними мітками, додатково перерозмічуються шістьма емоційними категоріями: *радість, страх, гнів, сум, відраза, подив.* Таким чином, фізична структура WordNet-Affect складається з шести файлів: *anger.txt, disgust.txt, fear.txt, joy.txt, sadness.txt, surprise.txt*, де кожен файл являє собою опис будь-якої категорії.

Вибір цих шести емоцій основане на психологічному дослідженні людини відносно не вербально виражений емоцій [6]. У Табл. 2 представлено данні про ці файли. Приклад структури сінсету словника WordNet-Affect:

a#01567385 grossly offensive to decency or morality; causing horror

ТАБЛ. 2

Структура словника WordNet-Affect

Файл	Кількість сінсетів	Кількість слів
joy.txt	128	318
anger.txt	20	72
disgust.txt	83	208
fear.txt	228	539
sadness.txt	124	309
surprise.txt	29	90
Усього	*612*	*1,536*

У таксономії WordNet-Affect є більше ніж 200 понять, пов'язаних різними типами відношень. На ресурсі «WordNet-Affect Taxonomy – SKOS version» є візуальне представлення таксономічного дерева WordNet-Affect [7].

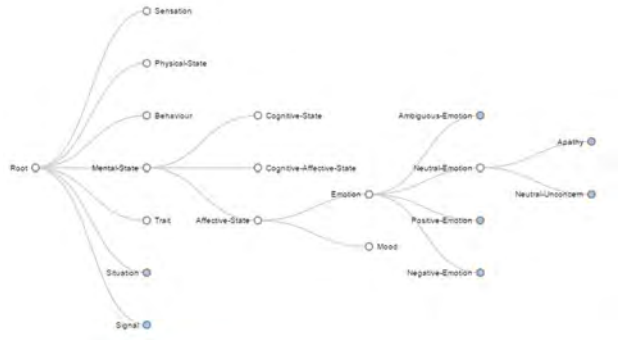

Рис. 1 Фрагмент таксономії WordNet-Affect.

IV.Процес створення тезаурусу WordNet-Affect для української мови

Для створення лексичного ресурсу на основі WordNet-Affect для української мови ми використовували готовий лексичний ресурс (http://lilu.fcim.utm.md/resourcesRoRuWNA_ru.html), який розробили дослідники з Технічного університету Молдови у роботі [8] для російської і румунської мов.

Основний процес створення української версії цього тезауруса полягав у перекладі. Проте складнощі все одно були.

WordNet-Affect складається з шістьох текстових файлів, що позначають такі емоції як: огида, страх, смуток, здивування, радість та гнів. У цих файлах міститься унікальний індекс емоції в якому перша буква позначає частину мови, наприклад, n – це noun (іменник), v – verb (дієслово) і так далі. Потім йде опис емоції англійською мовою. Після цього приклади слів або так звані сінсети (набір синонімів), що описують даний стан людини, записані в такому порядку: англійська-російська-румунська.

Приклад із файлу *disgust.txt* без української мови:

n#05577970 intense aversion repugnance repulsion horror отвращение oroare greață repulsie silă scârbă scîrbă dezgust Sentiment de dezgust, de scârbă, de neplăcere, de repulsie față de cineva sau de ceva

Перед нами ж стояло завдання додати до цього ряду ще й українську мову. Здебільшого переклад відбувався з російської мови на українську, але іноді було важко знайти підходящі еквіваленти і в такому випадку була використана ще й англійська мова.

Приклад із файлу *disgust.txt* з українською мовою:

n#05577970 intense aversion repugnance repulsion horror отвращение oroare greață repulsie silă scârbă scîrbă dezgust Sentiment de dezgust, de scârbă, de neplăcere, de repulsie față de cineva sau de ceva огида відраза огидливість обридливість обридження осоруга

Опишемо ресурси, які ми використовували для більш якісного та повного перекладу українською мовою. По-перше, переклад здійснювався повністю вручну без будь-яких допоміжних електронних програм. Кожне слово переводилося і перевірялося. Також ми намагалися знайти якомога більше синонімів до кожного слова, однак щоб даний синонім збігався за змістом з визначенням самої емоції. Отже, для перекладу слів ми використовували словники української мови [9], україно-англійські та англо-українські словники [10], словники синонімів [11] та деякі онлайн-словники [12, 13].

При перекладі з російської на українську у нас виникали деякі проблеми. Основна проблема була пов’язана з перекладом дієприкметників. Наприклад, такі слова як: *поклоняющийся, боготворящий,*

ценящий та інші. В українській мові дійсні дієприкметники теперішнього часу з суфіксами *-ущ- (-ющ-), -ащ- (-ящ-)* утворюються не від усіх дієслів і вживаються рідко. Ще рідше вони керують залежними словами. Російські дієприкметники, особливо утворені від дієслів з афіксом *-ся*, при перекладі на українську мову замінюють підрядними реченнями. Наприклад, *движущаяся колона – колона, що (яка) рухається*. Тобто слова з прикладу вище ми перевели як: *який (що) схіляється, який (що) боготворить, який (що) цінує*.

При перекладі з англійської мови, проблем не виникло. Однак це пов'язано з тим, що до англійської мови ми зверталися тільки у виняткових випадках, наприклад, коли хотіли доповнити синонімічний ряд. А так як в російській мові в більшості випадків було багато синонімів, це відбувалося досить рідко. Ось, наприклад, рядок взятий із файлу *surprise.txt*:

a#00405649 in a state of mental numbness especially as resulting from shock dazed stunned stupefied stupid
онемелый оцепенелый ошеломлённый потрясённый шокированный оцепеневший остолбеневший оглушённый uluit zăpăcit năuc năucit buimac uimit stupefiat

З цього прикладу можна побачити, що в англійській мові запропоновано чотири варіанти синонімів, а в російській мові вже вісім варіантів. Тобто з російської мови можна було взяти більше варіантів для поповнення синонімічного ряду. Табл. 3 та 4 представляє кількість сінсетів та слів для української та російської мов словника WordNet-Affect.

ТАБЛ. 3

Структура словника WordNet-Affect
для української мови

Файл	Кількість сінсетів	Кількість слів
joy.txt	206	1,340
anger.txt	116	524
disgust.txt	17	80
fear.txt	76	712
sadness.txt	96	1,118
surprise.txt	26	118
Усього	*537*	*3,892*

ТАБЛ. 4

Структура словника WordNet-Affect
для російської мови

Файл	Кількість сінсетів	Кількість слів
joy.txt	206	747
anger.txt	116	392
disgust.txt	17	73
fear.txt	76	330
sadness.txt	96	440
surprise.txt	26	112
Усього	*537*	*2,094*

Таким чином, було створено українську версію тезаурусу WordNet-Affect, який складається з 537 сінсетів та 3,892 слів.

Висновок

У цій статті наведено семантичний тезаурус WordNet-Affect, його структуру та особливості. Процес створення WordNet-Affect для української мови описано на основі лексичного ресурсу, розробленого науковцями Технічного університету Молдови для російської і румунської мов. У ході роботи було створено версію WordNet-Affect для української мови, який можна використовувати для автоматичної оцінки новин, продуктів, людей, організацій, країн та ін. Також цей семантичний тезаурус може бути використаний для розпізнавання і інтерпретації думки, кластеризації текстів, на основі полярності (позитивна, негативна, нейтральна) думок; сегментації текстів; прогнозування думок та ін.

Література

1. Батура Т. В. Основы обработки текстовой информации / Т. В. Батура, М. В. Чаринцева. — 2016. — 45 с.
2. Панг Б. Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval / Б. Панг, Л. Ли. — Москва : Вильямс, 2008. — 235 с.
3. Панг Б. Thumps up? Sentiment Classification using Machine Learning Techniques / Б. Панг, Л. Ли. — Москва : Вильямс, 2002. — 312 с.
4. Воронина И. Е. Анализ эмоциональной окраски сообщений в социальных сетях (на примере сети «вконтакте») / И. Е. Воронина, В. А. Гончаров // Вестник ВГУ, Серия: системный анализ и информационные технологи. — 2015. — № 4. — С. 151-158.
5. Strapparava C. WordNet-Affect: an Affective Extension of WordNet / C. Strapparava, A. Valitutti// The 4th International Conference on Language Resources and Evaluation (LREC 2004). —. Lisbon — Portugal, 2004. — P. 1083–1086.
6. Strapparava C. The affective weight of the lexicon / C. Strapparava, A. Valitutti, O. Stock // The 5th International Conference on Language Resources and Evaluation (LREC 2006). — Genoa. — Italy, 2006. — P. 474–481.
7. WordNet-Affect Taxonomy [Електронний ресурс]. – Режим доступу: https://www.gsi.dit.upm.es/ontologies/wnaffect/
8. Bobicev V. Emotions in words: developing a multilingual WordNet-Affect / V. Bobicev, V. Maxim, T. Prodan, N. Burciu, V. Angheluş // CICLing 2010, Ia□i. — Romania. — 2010. — P. 1-10.
9. Білецький-Носенко І.П. Словник української мови – Київ: Наук. думка, 1966. – 419 с.
10. Зубков М.Г. Сучасний англо-український словник – Х.: ВД «Школа», 2003. – 768 с.
11. Деркач П. М. Короткий словник синонімів української мови – Київ: Учбово-педагогічне видавництво «Радянська школа», 1960. – 210 с.

12. Онлайн словник «Академік» [Електронний ресурс]. – Режим доступу: http://dic.academic.ru/

13. Lingvo Online – бесплатный онлайн словарь [Електронний ресурс]. – Режим доступу: http://www.lingvo.ua/ru/Translate/uk-en

Електронний словник синонімів як засіб системного опису синонімічних відношень у лексичній системі української мови

Тетяна Грязнухіна

к. філол. н., старший науковий співробітник, старший науковий співробітник Українського мовно-інформаційного фонду НАН України, Україна, E-mail: ukritaldb@gmail.com

Тетяна Любченко

к. т. н., старший науковий співробітник Українського мовно-інформаційного фонду НАН України, Україна, E-mail: t.lyubch@gmail.com

The article is devoted to the questions of procedural determination of the synonyms with using the theory of the semantic states, to the questions of forming the Electronic dictionary of synonyms (EDS) of the Ukrainian language. EDS is oriented for the using it as an instrument for automatic semantic indexing of the text in the natural language processing systems.

Ключові слова — лексичні синоніми, синсет, електронний словник синонімів, семантичний стан, інтегрована лексикографічна система, комп'ютерні інформаційні технології.

I. Вступ

Синонімія належить до універсальних явищ, властивих усім природним мовам, що вирізняються з-поміж інших семіотичних систем своїм багато-однозначним і одно- багатозначним співвідношенням між позначуваним об'єктом і його позначальним знаком. Синоніми у мові є проявом другого із зазначених типів: різні знаки позначають той самий об'єкт.

Синонімічні відношення завжди були в центрі уваги лінгвістів-дослідників з лексичної семантики, основним положенням (твердженням) якої є декларування високого ступеня системної організації лексичної системи мови, сутність якої проявляється в різноманітних семантичних контекстах – синонімічних, антонімічних, паронімічних, гіпонімічних, без урахування яких здійснення повного опису семантики слова з опертям лише на його референтне значення неможливе. Насамперед це стосується багатозначних слів, доля яких в словнику мови становить біля 40%.

II. Об'єкт і завдання дослідження

Об'єктом даного дослідження є синонімічна підсистема української мови, тобто лексичні синоніми.

Синонімічні відношення порівняно з іншими вказаними вище парадигматичними відношеннями мають більш універсальний характер щодо вияву їх у межах різних лексико-граматичних класів слів. Якщо відношення гіпо- гіперонімії відіграє суттєву роль здебільшого в організації лексичного значення іменника, антонімічні відношення – прикметника, прислівника (градуальні, комплементарні) та дієслова (конверсиви), то синонімічні відношення є базовими в організації значення всіх частин мови.

Загострення уваги дослідників до явища синонімії на сучасному етапі розвитку лінгвістичної науки обумовлено також тією роллю, яку відіграє залучення інформації про синонімічні відношення при розв'язанні завдань автоматичного опрацювання мовної інформації, пов'язаних зі створенням таких нових засобів комунікації, як INTERNET, із завданням підвищення ефективності роботи систем автоматичної обробки текстів (АОТ). В інформаційних системах показник повноти і точності пошуку інформації значно збільшується із включенням до пошукового образу, не тільки ключових слів, але й їхніх синонімів. В системах автоматичного перекладу чітке розмежування перекладних еквівалентів та їхніх синонімічних відповідників призводить до зменшення кількості варіантів перекладу. Супровід перекладних еквівалентів синонімами в автоматичних перекладних словниках підвищує якість перекладу. Синонімічне індексування текстових корпусів забезпечує можливість проведення різних досліджень з питань функціонування синонімів у мовленні на репрезентативному текстовому матеріалі.

В будь-якій інтелектуальній системі АОТ основним джерелом, з якого здійснюється екстракція семантичної інформації, є словник. Є також розуміння того, що в єдиний словник неможливо вмістити всю різнобічну інформацію, необхідну для опису семантики мовної одиниці щодо її лексичного значення, парадигматичних і синтагматичних зв'язків, які є відображенням семантичних контекстів. Як показали результати досліджень з питань комп'ютерної лінгвістики, здійснюваних в Українському мовно-інформаційному фонді під керівництвом академіка НАНУ Широкова А.В., такий багатоаспектний комплексний опис семантики може