УДК 90.053

МЕТОДИКА ВИОКРЕМЛЕННЯ ТЕРМІНІВ ВЕТЕРИНАРНОЇ МЕДИЦИНИ З КОРПУСУ ВЕТЕРИНАРНИХ ТЕКСТІВ

METHODS OF VETERINARY TERMS RETRIVAL FROM THE CORPUS OF MEDICAL TEXTS

Рожков Ю.Г.,
orcid.org/0000-0002-6830-9130
асистент кафедри романо-германських мов і перекладу
Національного університету біоресурсів і природокористування України

Стаття присвячена створенню корпусу текстів ветеринарної тематики. Основна увага зосереджена навколо принципів виокремлення термінів із корпусу ветеринарних текстів. Для вищезазначених цілей було використано інструмент для класифікацій текстів. Для навчання класифікатора був підготовлений набір текстів, які були вручну класифіковані і завантажені в систему засобами сервісу класифікації. У роботі використаний гібридний метод: спочатку відбираються всі іменники, з яких потім за результатами статистичного аналізу з використанням інформації про семантику слів формується список можливих термінів. Для підвищення точності відбору в нашому дослідженні застосовувався кластерний аналіз англомовного тексту, який використовує адаптивний алгоритм Леска «за допомогою знаходження перетинів значень слів у WordNet». Результати застосування методу порівнювалися з результатами, отриманими під час використання частотного словника, частотного семантичного словника і позиційного методу. Таким чином, можна стверджувати, що послідовне застосування методів лінгвістичного і кількісного аналізу до спеціалізованого корпусу текстів дозволяє створити список кандидатів у терміни, різко скорочує роботу термінолога і дозволяє створювати реальні глосарії предметної області.

Ключові слова: корпусна лінгвістика, корпус текстів, терміни, термінологія, ветеринарна медицина.

Статья посвящена созданию корпуса текстов ветеринарной тематики. Основное внимание сосредоточено вокруг принципов выделения терминов из корпуса ветеринарных текстов. Для вышеуказанных целей были использованы инструменты для классификации текстов. Для обучения классификатора был подготовлен набор текстов, который был вручную классифицирован и загружен в систему средствами сервиса классификации. В работе использован гибридный метод: сначала отбираются все существительные, из которых затем по результатам статистического анализа с использованием информации о семантике слов формируется список возможных терминов. Для повышения точности отбора в нашем исследовании применялся кластерный анализ англоязычного текста, который использует адаптивный алгоритм Леска «посредством нахождения пересечений значений слов в WordNet». Результаты применения метода сравнивались с результатами, полученными при использовании частотного словаря, частотного семантического словаря и позиционного метода. Таким образом, можно утверждать, что последовательное применение методов лингвистического и количественного анализа в специализированном корпусе текстов позволяет создать список кандидатов в термины, резко сокращает работу терминологов и позволяет создавать реальные глоссарии предметной области.

Ключевые слова: корпусная лингвистика, корпус текстов, терминология, ветеринарная медицина.

The article is dedicated to creation of veterinary oriented texts corpus. The attention is focused on the principles of finding and retrieval of terms from the veterinary oriented texts corpus. For the targets mentioned above we used various instruments for texts classification. In order to teach classificator the collection of texts, which was selected manually and downloaded into the system by means of classification services, was prepared. We used the hybrid method: at first we selected all the nouns which eventually form the list of plausible term candidates. To increase the precision of selection cluster analysis of English text which uses adoptive pattern of Lesk was implemented. The results of implemented method were compared to the results received while using frequency dictionary, frequency semantic dictionary and positional method. Thus one may claim that sequential usage of linguistic and quantitative analysis in specialized texts corpus allows creating the list of term candidates, rapidly minimizes the work of terminologists and creates real glossaries of subject area.

Key words: corpus linguistics, corpus of texts, terminology, veterinary medicine.

Постановка проблеми. Сьогодні різного роду задачі трактуються за допомогою когнітивного підходу, який дозволяє детальніше дослідити мовні явища.

Оскільки медичні та ветеринарні тексти подають основну інформацію через терміни, їх вивчення допоможе зрозуміти та розтлумачити процеси, що відбуваються всередині термінологічної системи. Когнітивний підхід є одним із головних методів дослідження термінологічних одиниць, він дає змогу пізнати рівні функціонування термінів, а також сприяє процесу пізнання в цілому.

Аналіз останніх досліджень і публікацій. У наш час із дослідженням розумової діяльності людини пов'язують рішення різноманітних питань мовознавства. Термінологія, де кожен термін має чітку, точну структуру знання, являє собою цікавий об'єкт для когнітивного мовознавства [1, с. 61].

Когнітивний напрямок термінознавства в Україні розробляється М.М. Полюжиним, С.А. Жаботинською, О.П. Воробйовою та іншими.

Виклад основного матеріалу. Корпусна лінгвістика сьогодні часто розуміється як відносно новий підхід у мовознавстві, що пов'язано з емпіричним вивченням мови «реального життя» за допомогою комп'ютерів та електронних корпусів. У першу чергу, «корпус» – це певний збірник письмових або розмовних текстів.

Однак коли термін використовується в контексті сучасної лінгвістики, він має тенденцію нести ряд конотацій, серед них – машиночитабельна форма, вибірка і репрезентативність, кінцеві розміри, а також ідея, що корпус являє собою стандартизовану базу даних мови, яку він представляє. Хоча мовознавство ділиться на багато областей досліджень залежно від комплексів дослідницьких питань, корпусна лінгвістика, по суті, веде себе діаметрально протилежно: вона пропонує набір методів, які можуть бути використані в дослідженні великої кількості різних дослідницьких проблем [9, с. 75].

Як методологія сучасне корпусне мовознавство тісно пов'язане з історією лінгвістики як емпіричної науки. Багато методик, які використовуються в корпусній лінгвістиці, набагато старші за комп'ютери: багато з них своїм корінням уходять у традиції пізнього вісімнадцятого і дев'ятнадцятого століття, коли лінгвістика була вперше визнана як «справжня», або емпірична наука.

Одним з основних факторів виникнення сучасного мовознавства є порівняльне та історичне мовознавство, причиною чого є, звичайно, те, що дослідники в області історичного мовознавства завжди використовували тексти або текстові колекції як матеріал для своїх досліджень [2, с. 111].

Багато методик, які були розроблені в дев'ятнадцятому столітті для реконструкції старих мов або визначення відносин між ними, досі використовуються. В індоєвропейської традиції вивчення зміни мови та її реконструкції залежали від ранніх текстів або корпусів [1, с. 134]

Якоб Грімм, а пізніше і неограматики підтримували свої твердження про історію та граматику мов шляхом цитування текстових даних. Неограматики висловлювали думку, що вивчення сучасної мови через діалекти (а не тільки вивчення ранніх текстів) є актуальним питанням.

Багато ідей та методів, розроблених дослідниками дев'ятнадцятого століття, сьогодні були адаптовані та набули подальшого розвитку в сучасній корпусній лінгвістиці. Сьогодні існує великий інтерес у складанні історичних корпусів; історичні корпуси були серед перших корпусів, доступних в електронному вигляді (наприклад, новаторська робота Роберто Буза, присвячена творам Томаса Аквінського, та зразки Августинської прози Луї Міліча) [3, с. 69].

Поява електронних текстів дала змогу зібрати величезні суми даних порівняно швидко. Це, у свою чергу, дало змогу науковцям скористатися перевагами статистичних методів у лінгвістичному аналізі та розробляти і розвивати нові сучасні моделі та інструменти для їх досліджень. Сьогодні математично складні моделі мовної зміни можуть обчислюватися за допомогою даних, отриманих з електронних корпусів.

Граматики дев'ятнадцятого століття ілюстрували свої твердження, зроблені за допомогою прикладів із творів визнаних авторів. Наприклад, Герман Пауль у його **Prinzipien der Sprachgeschichte** використовував німецьку «класику», щоб проілюструвати практично будь-яке твердження, чи то у фонології, в морфології або в синтаксисі [4, с. 145].

Сьогодні укладачі граматичних підручників можуть також застосовувати корпусний підхід, але корпуси, які вони використовують, включають у себе не тільки класику, а й усі види текстів.

У лексикографії, наприклад, Оксфордський англійський словник і багато словників мертвих мов дають цитати з текстів, що містять слово, яке розглядається в контексті. У сучасному корпусному мовознавстві даний метод реалізується у формі конкордансів (ключове слово в контексті – KWIC). Навіть незважаючи на те, що комп'ютери спрощують пошук і класифікацію статистики для виокремлення колокацій і цікавих шаблонів для кожного слова, основоположні методи використання текстових корпусів все ще дуже схожі на ті, які використовували ранні лексикографи і філологи, що не мали доступу до комп'ютерних технологій.

Традиційні підручники з граматики вищої школи, як правило, містять відредаговані приклади використання мови. У довгостроковій перспективі вони можуть забезпечити лише обмежену підтримку студентам, які рано чи пізно стикаються з автентичними мовними даними у своїх завданнях. У цьому відношенні важливу роль у мові відіграють корпуси як джерела емпіричних даних.

У багатьох дисциплінах зараз розробляються стандартні онтології предметних областей, які призначені для спільного використання експертами і автоматичними системами обробки інформації. Процес створення онтології характеризується великим об'ємом, оскільки необхідно адекватно і максимально повно описати кожен концепт (термін), що входить у неї, із зазначенням усіх можливих зв'язків з іншими концептами.

У нашому дослідженні аналізується початковий етап побудови онтології предметної області – автоматичне формування списку термінів. В якості вихідного матеріалу були взяті статті з Black's Veterinary Dictionary. Представлені результати кластеризації корпусу визначень ветеринарних понять із подальшим виділенням термінів для кожного кластера. Обсяг корпусу був обмежений розмірами, що дозволяє вести ручну експертну обробку для контролю якості роботи автоматики [5, с. 549].

З бурхливим зростанням кількості оброблюваної інформації останні десятиліття потреба в розвитку методів та інструментів комп'ютерної лінгвістики тільки збільшується. Одним із завдань комп'ютерної лінгвістики є автоматична класифікація текстів, тобто віднесення тексту до тієї чи іншої області або її підмножини на основі деякого алгоритму з певною ймовірністю.

Частина алгоритмів використовує для цього тільки дані, отримані безпосередньо з даного тексту. Такі алгоритми мають невисоку точність і часто не відповідають рішенням задачі класифікації людиною, частина алгоритмів використовує додаткову інформацію (навчальні вибірки текстів, словники предметних областей, списки слів-ознак і т. д.), що вимагає підготовки додаткових даних.

Починаючи з ранніх етапів розвитку, окремий суб'єкт пізнання розвиває своє уявлення про навколишній світ, набуваючи нові знання, і з кожним разом усе краще справляється із завданням класифікації тексту, а будучи фахівцем із будь-якої більш вузької області, може максимально точно класифікувати текст. Тому будь-яка комп'ютерна система класифікації повинна бути самонавчальною, незалежно від використовуваного алгоритму класифікації, з кожним разом вирішуючи завдання все точніше, використовуючи весь накопичений за час роботи досвід [6, с. 345].

У комплексі інструментів автоматизованого аналізу текстів реалізовані інструменти аналізу і дослідження текстів на етапах морфологічного, синтаксичного аналізу, із застосуванням статистичних методів. Крім того, присутній засіб дослідження отриманих результатів на наступному аналітичному рівні. На основі інструментів комплексу створені сервіси вирішення завдань виділення ключових слів, статистичного аналізу, класифікації, представлені на порталі «Автоматизований аналіз тексту» [7, с. 61].

Сервіс класифікації текстів має два режими роботи: режим аналізу і режим навчання, в основі якого – робота з ключовими словами, отриманими в результаті розрахунку частоти вживання слів у тексті із застосуванням морфологічного аналізу і засобів аналітичної обробки. Упровадження інших методів класифікації дозволить поліпшити точність результатів, але аналіз результатів класифікації показав істотне збільшення якості одержуваних результатів після навчання класифікатора в рамках однієї предметної області.

Для навчання класифікатора був підготовлений набір текстів, які були вручну класифіковані і завантажені в систему засобами сервісу класифікації. Для навчання класифікатора на великому наборі предметних областей попереду стоїть завдання повної автоматизації цього процесу, що вимагає наявності розміченого корпусу текстів.

Існує безліч класифікацій корпусів текстів: за способом їх побудови (статичні і динамічні, одномовні і багатомовні, розмічені і немарковані і т. д.), поширення (вільно або частково доступні, закриті), призначенням і т.д. [7, с. 158].

Залежно від розв'язуваної задачі виникає необхідність у наявності різної інформації про тексти в корпусі: морфологічну, синтаксичну, семантичну. Наприклад, для вирішення завдання визначення тональності тексту необхідні вказівка частин мови слів, розмітка пропозицій, семантична розмітка та інше. Для навчання системи класифікації текстів потрібна розмітка корпусу текстів відповідно до обраних ознак класифікування (стиль тексту, жанр, автор, тематика, дата написання і т.д.). Крім того, корпус текстів повинен бути досить великим, щоб забезпечувати репрезентативність вибірки та прийнятну якість навчання класифікатора, в тому числі у випадку збільшення кількості одиниць розбиття (бінарна, багатокритеріальна або фасетна класифікації).

Існують онлайн-варіанти корпусів, але вони не дуже придатні для автоматичного навчання класифікатора у зв'язку з обмеженням списку результатів, що повертаються, і закритістю вихідних корпусів текстів.

Одним із популярних напрямків у комп'ютерній лінгвістиці є визначення тональності тексту або відкликання про що-небудь (наприклад), велика частина відомих алгоритмів вирішення цього завдання спирається на методи машинного навчання, і, як наслідок, під час застосування даних методів потрібна підготовка корпусів текстів. Часто маючи достатній обсяг, вони призначені тільки для навчання класифікатора для визначення тональності текстів на основі ручного аналізу добірки текстів.

Таким чином, необхідне розроблення корпусу текстів із розміткою, придатною для навчання класифікатора текстів незалежно від способу його реалізації (використання статистичних методів комп'ютерної лінгвістики, машинного навчання,

нейронних мереж). Такий корпус повинен бути універсальним, придатним для навчання класифікаторів із різним поповнюваним набором ознак і з можливістю створення на його основі субкорпусів для вирішення того чи іншого завдання.

Відбір термінів для включення до онтології зазвичай здійснюється за допомогою лінгвістичного і статистичного аналізу [8, с. 72–78]. У роботі використаний гібридний метод: спочатку відбираються всі іменники, з яких потім за результатами статистичного аналізу з використанням інформації про семантику слів формується список можливих термінів.

Для підвищення точності відбору в нашому дослідженні застосовувався кластерний аналіз англомовного тексту, який використовує адаптивний алгоритм Леска «за допомогою знаходження перетинів значень слів у WordNet». Результати застосування методу порівнювалися з результатами, отриманими під час використанні частотного словника, частотного семантичного словника і позиційного методу. Варто відзначити, що в попередніх дослідженнях спочатку показники повноти і точності у визначенні ключових слів за частотним словником були досить високими. Схожа методика використовується і в нашій роботі. Однак метою роботи було вилучення ключових слів, які характеризують кожен текст із корпусу, а мета нашого дослідження – витяг термінів, що характеризують конкретну галузь знань.

Для контролю якості роботи системи попередньо вручну було виділено 462 однослівних терміна. Ми зосередилися саме на однослівних термінах, тому що їх автоматичне виділення представляє найбільші труднощі. Двослівні терміни в даних текстах мають чітко виражену структуру: прикметник + іменник, розташовані контактно, і можуть бути знайдені стандартними частотними методами. Обраний тип тексту (словник) має свої особливості. З одного боку, всі словникові статті побудовані за одним шаблоном: заголовок статті і дефініція, що складається, як правило, з гіперонімів і додаткової інформації.

При цьому терміни предметної області можуть зустрічатися в обох частинах словникової статті. Як виявилося, в заголовну частину входить тільки 59% термінів. Така структурованість полегшує оброблення тексту. З іншого боку, для текстів наукової спрямованості типовою є ситуація, за якої одні терміни предметної області зустрічаються дуже часто (БАНК – 826 входжень, 3% від усіх неслужбових слів корпусу), а інші – мають лише поодинокі входження (анаплазма, цистоплазмоз та ін.).

У процесі використання стандартних методів відбору лексем (типу TF-IDF, LDA) для подальшого оброблення відкидаються як ті, так і інші. У той же час серед слів із середньою частотністю більшу частину складають слова загальної лексики. Наприклад, термін СКАЛЬПЕЛЬ з абсолютною частотою входжень 45 в уже згадуваному корпусі має за частотою іменників 56-й ранг з 98. У першому стовпчику табл. 1, який було збудовано за частотним словником, наведені 19 іменників з околиці слова СКАЛЬПЕЛЬ в інтервалі з 50 рангу (частота 51) по 62 ранг (частота 39).

Висновки. У результаті такого аналізу всі виокремленні кандидати в терміни об'єднуються для ручної перевірки і отримання остаточного списку термінів. Таким чином, можна стверджувати, що послідовне застосування методів лінгвістичного і кількісного аналізу до спеціалізованого корпусу текстів дозволяє створити список кандидатів у терміни, різко скорочує роботу термінолога і дозволяє створювати реальні глосарії предметної області.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ:

1. Friedrich J. (ed.). Kleinasiatische Sprachdenkmäler. *Walter de Gruyter*, 2013. T. 163.
2. Mair C., Hundt M. (ed.). Corpus linguistics and linguistic theory. *Rodopi*. 2000. № 1999.
3. Winter T. N. Roberto Busa, SJ, and the invention of the machine-generated concordance. Faculty Publications, Classics and Religious Studies Department. 1999. C. 70.
4. Paul H. Prinzipien der sprachgeschichte. *Walter de Gruyter*. 2010. T. 6.
5. West G. P. Black's veterinary dictionary. *Rowman & Littlefield*, 1998.
6. Biber D. et al. Corpus linguistics: Investigating language structure and use. *Cambridge University Press*. 1998.
7. Stubbs M. Text and corpus analysis: Computer-assisted studies of language and culture. Oxford : Blackwell, 1996. C. 158.
8. Гаврилова М.В. Лингвистический анализ политического дискурса. *Политический анализ*. 2003. №. 3. С. 72–78.
9. Рожков Ю.Г. Лінгвоконітивний підхід до вивчення термінології ветеринарної медицини. *Scientific Notes of Ostroh Academy National University: Philology Series.* 2017. №. 66. С. 75–76.