УДК 801.009

В.В. Балабін, канд. філол. наук, проф.,
І.В. Замаруєва, д-р техн. наук, проф.,
КНУ імені Тараса Шевченка

РОЗРОБКА ФОРМАЛЬНОЇ МОДЕЛІ ПРЕДСТАВЛЕННЯ СИНТАГМАТИЧНИХ І ПАРАДИГМАТИЧНИХ ВІДНОШЕНЬ В ПРЕДМЕТНІЙ ОБЛАСТІ ДЛЯ РІЗНОМОВНИХ ТЕКСТІВ

У статті запропоновано підхід до автоматизації формування баз знань про предметну область на основі корпуса різномовних текстів заданої тематики. Розглянуто й обґрунтовано структуру тезауруса. Запропоновано процедуру наповнення тезауруса безпосередньо за корпусом текстів, визначені вимоги щодо повноти та інформативності корпуса текстів.

Ключові слова: корпус різномовних текстів, база знань про предметну область, тезаурус, парадигматичні відношення, синтагматичні відношення.

Актуальність дослідження. Проблема автоматизації розуміння природно-мовних текстів (ПМТ) нерозривно пов'язана із побудовою баз знань, оскільки саме на підставі певних знань про навколишній світ (предметну область) людина здатна розуміти інші тексти із заданої тематики. Існуючі підходи до побудови баз знань (реляційні, предикатні, продукційні моделі) сьогодні не орієнтовані на автоматизоване вилучення знань про предметну область безпосередньо із тексту. В той же час збільшення електронних обсягів ПМТ і задач щодо їх опрацювання потребують оперативного поповнення відповідних словників – тезаурусів. До таких задач і відносяться системи машинного перекладу, для яких тезаурус значно підвищив би якість перекладу. Пропонується підхід до автоматизації побудови бази знань та процедуру наповнення тезауруса безпосередньо за корпусом текстів**.**

Постановка проблеми дослідження. Основними компонентами знань з точки зору їх формалізованого подання є поняття, відношення між ними, характеристики понять і відношень, а також модальності цих характеристик. Отже, обробка вхідного тексту має бути спрямованою на виявлення (розпізнавання) в тексті основних компонент знань і встановлення логіко-семантичних відношень між ними з метою формування поняттєвої структури змісту вхідного тексту.

До формалізованого представлення знань пред'являються наступні вимоги:

- по-перше, воно має бути подано в такому вигляді, який забезпечить можливість коректної логіко-семантичної обробки знань (в умовах багатозначності і невизначеності текстових одиниць);

- по-друге, воно має містити всю необхідну інформацію для забезпечення адекватного перекладу, тобто максимально повно зберігати текстове представлення елементів знань.

З урахуванням цих вимог в якості формалізованого подання знань вибрана поняттєва структура (ПС) змісту природно-мовного тексту (ПМТ). Вона являє собою ієрархічну структуру, на верхньому рівні якої знаходяться найбільш загальні поняття і відношення між ними, кожний нижній рівень представляється поняттями і відношеннями, які конкретизують відповідні поняття і відношення найближчого вищого рівня. Сформована таким чином ПС містить всю необхідну інформацію для вирішення прикладних задач машинного перекладу. Можливість її формування визначається наявністю відповідних знань в тезаурусі системи.

Особливості ПС полягають в наступному: її подання є гібридним і поєднує в собі властивості семантичних мереж і предикатних моделей (в якості вершин мережі виступають предикати); з метою уніфікації подання відношень, які в тексті можуть мати різну кількість аргументів, в ПС використовуються тільки одно- та двохмісні предикати, для чого розроблено метод декомпозиції n-місцевих предикатів і предикатів вищих порядків на двохмісні предикати першого порядку. Для відображення рольових відношень введене поняття неявних предикатів; з метою зберігання виразових засобів природно-мовного текстового подання, введені спеціальні засоби – префікси і постфікси предикатів і понять, логіко-лінгвістичні зв'язки, анафоричні посилання тощо.

Змістове наповнення ПС суттєво залежить від формалізації конкретних відношень, понять та їх характеристик, притаманних саме визначеній предметній області (ПО). ПМТ, що містить фрагменти знань про ПО визначається парадигматичними і синтагматичними відношеннями. Парадигматичні відношення ідентифікують системні зв'язки між поняттями в ПО. Такі відношення, як правило, не відносять до конкретного тексту, оскільки там не реалізуються. Парадигматичні відношення фактично характеризують професійну компетентність фахівця.

Викладення результатів дослідження.

Процес побудови моделі знань відбувається у декілька етапів. На першому етапі фахівець (експерт) укладає систему базових понять в заданій ПО з відповідними прагматичними відношеннями. Таку систему прийнято представляти у вигляді тезауруса. Призначення даного тезауруса – представити парадигматичні (так звані «вертикальні») відношення між базовими поняттями, що існують в ПО, і які не залежать від їх контекстного вживання.

Незалежно від ПО можна виділити парадигматичні відношення, які характеризують систему, а не конкретну ПО. Серед таких відношень виділяють такі: *частина, ціле*, *рід, вид*, *синонім*, *антонім, асоціативні відношення*. Асоціативні відношення відносяться до слабко формалізованих відношень можуть мати різний прагматичний зміст в залежності від ПО.

Крім того, існують системні відношення, які характеризують тільки конкретну ПО. Наприклад, ПО: *міжнародне право* – характеризується такими системними відношеннями, як *суб'єкт* і *об'єкт*, які не мають нічого спільного з категоріями *суб'єкт* і *об'єкт* під час семантичного аналізу тексту.

Для кожного відношення в тезаурусі задаються окремі поля. Формат представлення тезауруса показаний в табл.1.

Таблиця 1

Структура тезауруса

Код	**Дескриптор**	**рід**	**вид**	**ціле**	**частка**	**синонім**	**антонім**		**асоціації**		
1	2	3	4	5	6	7	8	9	10	11	12

Поля 1-8, як правило, є обов'язковими і мають однакове прагматичне значення, поля 9-12 мають різні прагматичні значення і взагалі можуть бути відсутніми.

Поле 1 визначає унікальний код відповідного поняття. Якщо ПО добре структурована, то код може містити закодовану семантичну інформацію про місце відповідного поняття в системі. Наприклад: 1.1.1. – означає, що поняття знаходиться на третьому рівні ієрархії в системі. Якщо ПО слабко формалізована, то доцільно ставити в якості коду порядковий номер поняття в тезаурусі.

Поле 2 містить саме поняття, яке представляється словом або словосполученням у початковій формі, його прийнято називати дескриптором.

Поле 3 містить родовий дескриптор для дескриптора, заданого в полі 2, якщо визначений дескриптор сам є родовим поняттям, то поле залишається незаповненим.

Поле 4 містить всі видові дескриптори для дескриптора, заданого в полі 2 (наприклад, для дескриптора: *меблі* визначаються його видові дескриптори: *офісні меблі*, *кухонні меблі* тощо), якщо визначений дескриптор є видовим поняттям найнижчого рівня, то поле залишається незаповненим.

Інші поля заповнюються аналогічним чином. Слід зазначити, що для конкретного поняття наповненість всіх полів не обов'язкова. Приклад заповнення представлений в табл. 2.

Таблиця 2

Приклад формування словникової статті в тезаурусі для поняття «національна безпека»

Код	Дескриптор	Ціле	Частина	Синонім	Антонім	Суб'єкт	Об'єкт
1.	Національна безпека		Безпека у воєнній сфері / Воєнна безпека	Безпека держави	Війна/ Воєнні дії/ Збройний конфлікт	Президент/ Рада національної безпеки і оборони / Збройні сили/	Людина/ Громадянин/ Суспільство/ Держава

З таблиці 2 видно, що поля 4-8 мають суто прагматичне наповнення, так визначаються не всі складові національної безпеки, а лише ті, які є актуальними для текстів військової тематики (в Законі України «Про основи національної безпеки України» визначається 10 складових), відношення *антонім* також має прагматичне наповнення, оскільки *небезпека*, як найбільш загальний антонім не розкриває його прагматичну сутність.

Даний тезаурус має подвійне призначення: по-перше, він дозволяє на етапі інтерпретації добирати коректні з точки зору ПО синоніми, якщо у перекладному словнику словника стаття містить декілька перекладних інваріантів, по-друге, словникова стаття в тезаурусі – це фактично готовий пошуковий образ запиту для формування корпусу текстів відповідної тематики, який є необхідної умовою побудови моделі синтагматичних відношень в ПО.

Процес формування синтагматичної моделі продемонструємо на прикладі. Нехай за нашим запитом ми набрали декілька фрагментів різномовних текстів (рис.1.).

Українська мова

Національна безпека – захищеність життєво важливих інтересів **людини** і **громадянина**, **суспільства** і **держави**, за якої забезпечується сталий розвиток **суспільства**, своєчасне виявлення, запобігання і нейтралізація реальних та потенційних загроз національним інтересам ……

Англійська мова

National security and defense can be understood as preparedness for **military action**, protection of resources considered critical to the function of a nation to protect a country from attack or subversion.
There are different **government agencies** concerned with **national security**, e.g., the **National Security Council** (NSC), the Central Intelligence Agency (CIA), the Federal Bureau of Investigation (FBI) – in the United States of America, ……

Російська мова

Настоящая Стратегия является базовым документом по планированию развития системы обеспечения **национальной безопасности** Российской Федерации, в котором излагаются порядок действий и меры по обеспечению **национальной безопасности**. Она является основой для конструктивного **взаимодействия органов государственной власти,** организаций и общественных объединений для защиты национальных интересов Российской Федерации и обеспечения безопасности личности, **общества** и **государства. …**

Рис. 1. Приклад фрагменту текстів за пошуковим образом «національна безпека»

Синтагматичні відношення в ПО визначають закономірності сполучуваності понять і відношень в певному тексті. Синтагматичну модель ПО можна побудувати лише на підставі вивчення навчальної вибірки текстів заданої

тематичної спрямованості, якщо мова йде про машинний переклад, то тематична вибірка має містити різномовні тексти. Тому, на другому етапі за усіма дескрипторами, що увійшли до тезауруса (парадигматичної моделі), формується корпус різномовних текстів відповідно до заданої тематики. Слід зазначити, що пошуковий образ можна формувати автоматично (і сьогодні розроблені відповідні програмні засоби) або вручну – для цього потрібно для всіх дескрипторів словникової статті тезауруса надати перекладні еквіваленти (в нашому випадку англійські й російські).

Складність побудови синтагматичної моделі знань про ПО за різномовними текстами полягає в тому, що відображення картини світу (ПО) засобами мови у різних народів не співпадає. Це пов'язане як із різними професійними поглядами на сутність явищ, фактів, способом доведення, так і об'єктивною різницею в самій картині світу.

На рис. 1 жирним шрифтом виділені ті лексеми, які були присутні в пошуковому запиті. Насиченість лексем із пошукового образу запиту свідчить про те, що відібраний текст, придатний для побудови синтагматичної моделі. І навпаки, якщо на задану довжину тексту лексеми із пошукового образу запиту зустрічаються з низькою частотою, то текст вважається не придатним для побудови синтагматичної моделі. Вимоги щодо насиченості тексту по відношенню довжина/частота визначає дослідник. Процес формування синтагматичної моделі відбувається окремо за кожним визначеним дескриптором.

Після опрацювання навчальної вибірки формуються синтагматичні відношення для кожного заданого дескриптора. Синтагматичні відношення для дескриптора «національна безпека», які проявилися в текстах з рис.1 представлені на рис. 2.

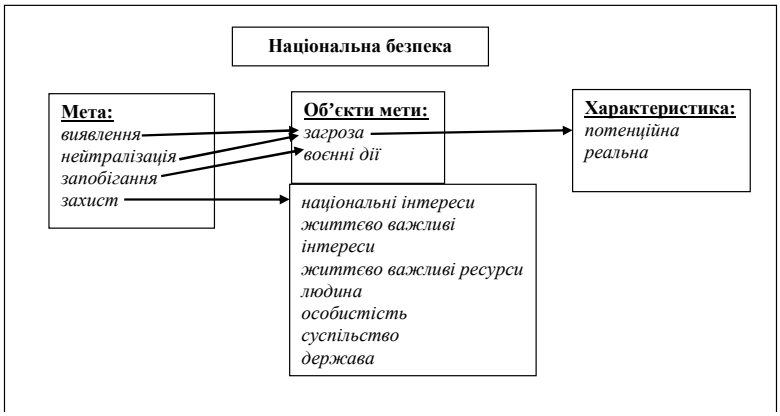

Рис. 2. Фрагмент синтагматичної моделі поняття "національна безпека"

Після побудови синтагматичної моделі корегується тезаурус. Такий крок пов'язаний з тим, що класифікація знань про ПО в різних мовах, як правило, не співпадають. Так, *об'єктами* захисту в Україні є *людина*, *громадянин*, *суспільство*, *держава*, а в РФ – *особистість, суспільство*, *держава.*

На останньому етапі парадигматична і синтагматична моделі об'єднуються в модель знань про ПО. При цьому синтагматичні відношення відбивають горизонтальні відношення семантичної мережі (вузлами якої є елементарні предикати), парадигматичні відношення – вертикальні відношення (які визначають ієрархію понять в ПО та інші системні відношення).

Висновки. Фрагменти знань, які описуються в ПМТ, відбивають стан фахового (або, в загальному випадку, логіко-семантичного) проникнення в ПО, а не певної природної мови. З урахуванням цього розроблена модель знань про ПО, яка являє собою інтеграцію парадигматичних (системних) і синтагматичних (текстових) відношень.

Запропонована модель представлення знань про ПО дозволяє усувати синтаксичну омонімію при автоматичному перекладі, а також у разі появи у тексті нового поняття, відсутнього у перекладному словнику, автоматично добрати контекстний синонім до нового слова.

Надійшла до редколегії 12.02.13

В.В. Балабин, канд. филол. наук, проф.,
И.В. Замаруева, д-р техн. наук, проф.,
КНУ имени Тараса Шевченко

РАЗРАБОТКА ФОРМАЛЬНОЙ МОДЕЛИ ПРЕДСТАВЛЕНИЯ СИНТАГМАТИЧЕСКИХ И ПАРАДИГМАТИЧЕСКИХ ОТНОШЕНИЙ В ПРЕДМЕТНОЙ ОБЛАСТИ ДЛЯ РАЗНОЯЗЫЧНЫХ ТЕКСТОВ

В статье предложен подход к автоматизации формирования баз знаний о предметной области на основе корпуса разноязычных текстов заданной тематики. Рассмотрена и обоснована структура тезауруса. Предложена процедура наполнения тезауруса непосредственно по корпусу текстов, сформулированы требования к полноте и информативности корпуса текстов.

Ключевые слова: корпус разноязычных текстов, база знаний о предметной области, тезаурус, парадигматические отношения, синтагматические отношения.

V.V. Balabin, PhD, Professor,
I.V. Zamaruieva, Doctor, Professor,
Taras Shevchenko National University of Kyiv

DEVELOPING A FORMAL MODEL FOR REPRESENTATION OF SYNTAGMATIC AND PARADIGMATIC RELATIONS IN THE SUBJECT AREA OF MULTI-LINGUAL TEXTS

The paper proposes an approach to the automated knowledge base formation on the subject area, based on the body of multilingual texts of given topic. The structure of the thesaurus was considered and justified. The procedure for filling thesaurus directly behind the body text was suggested, and the requirements of completeness and informativeness of the body text were defined.

Keywords: a body of multilingual texts; the knowledge base on the subject area; thesaurus; paradigmatic relations; syntagmatic relations.