УДК 81.302.2

С.В. Лєнков,
доктор технічних наук, професор,
І.В. Замаруєва,
доктор технічних наук, професор,
І.В. Пампуха,
кандидат технічних наук, доцент,
Л.В. Охрамович

МОДЕЛЬ ПРЕДСТАВЛЕННЯ ЗНАНЬ ДЛЯ БАГАТОМОВНОЇ СИСТЕМИ МАШИННОГО ПЕРЕКЛАДУ

У статті розглядається питання побудови моделі представлення знань про предметну галузь. На відміну від моделі знань про навколишній світ, запропонована модель призначена для прагматичного аналізу природно-мовного тексту, оскільки вона формалізує професійно орієнтовані знання. Необхідність введення цієї моделі до системи машинного перекладу обумовлена тим, що адекватність перекладу спеціальних (професійно орієнтованих) текстів, у першу чергу, залежить від фахової компетентності перекладача.

Ключові слова: *системи машинного перекладу, професійно орієнтовані знання, моделі представлення знань, адекватність перекладу.*

В статье рассмотрены вопросы построения модели представления знаний о предметной области. В отличие от моделей знаний об окружающем мире, предложенная модель предназначена для прагматического анализа природно-языковых текстов, поскольку она формализует профессионально ориентированные знания. Необходимость введения этой модели в систему машинного перевода обусловлена тем, что адекватность перевода специальных (профессионально ориентированных) текстов, в первую очередь, зависит от профессиональной компетентности переводчика.

Ключевые слова: *системы машинного перевода, профессионально ориентированные знания, модели представления знаний, адекватность перевода.*

Paper addresses the issue of model building of knowledge support about the subject area. Unlike the model of knowledge about the world proposed model is designed for the pragmatic analysis of natural language text (NLT) as it formalizes professionally oriented knowledge. The need to introduce this model to a machine translation system is grounded with the fact that the adequacy of the translation of special (professionally oriented) texts primarily depends on the professional competence of an interpreter.

Keywords: *machine translation systems, professional-reference knowledges, models of representation of knowledge, adequacy of translation.*

Особливості аналізу природномовного тексту (ПМТ) для знання-орієнтованих систем машинного перекладу (СМП) визначаються спрямованістю на формування понятійної структури, тобто на автоматичний витяг знань з різномовних текстів та їх прагматичну інтерпретацію в термінах предметної галузі. При цьому

текст розглядається як об'єкт різних рівнів аналізу: як знакова система, як граматична система і як система знань про світ (проблемну область). Кожний рівень має свої особливості, свої засоби представлення, а отже, припускає наявність специфічних методів обробки. Таким чином, СМП або взагалі не враховується професійна спрямованість тексту, або СМП є дуже вузькоспеціалізованою, що унеможливлює її застосування навіть у дуже близькій предметній галузі (ПГ), не кажучи вже про адекватний переклад професійно орієнтованих текстів, які включають декілька суміжних ПГ [1–5].

До формалізованого представлення знань пред'являються такі вимоги [1; 5]: по-перше, воно має бути подано в такому вигляді, який забезпечить можливість коректної логіко-семантичної обробки знань (в мовах 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 як формалізоване подання знань обрана понятійна структура (ПС) змісту ПМТ. Вона становить ієрархічну структуру, на верхньому рівні якої знаходяться найбільш загальні поняття і відношення між ними, кожний нижчий рівень представлений поняттями і відношеннями, які конкретизують відповідні поняття і відносини найближчого вищого рівня. Іншими словами, верхній рівень ПС відповідає найбільш загальному опису змісту тексту, нижчі її рівні відповідають рівням конкретизації цього опису. Кожне поняття і вид відносин в ПС супроводжується характеристиками, які відображають їхні властивості (понять і відносин), модальності та інші аспекти; лінгвістичною інформацією, яка характеризує мовні засоби їх відображення у вхідному тексті; семантичною інформацією, яка відбиває їх роль та інші характеристики (наприклад: *об'єкт, суб'єкт, тип відношення, напрямок дії тощо*). Сформована таким чином ПС містить всю необхідну інформацію для вирішення прикладних задач машинного перекладу. Можливість її формування визначається наявністю відповідних знань в тезаурусі системи [1–3].

ПС, яка задовольняє сформульовані вимоги і містить всю необхідну інформацію як для подальшої її логіко-семантичної обробки, так і для синтезу опису ПС або її фрагментів природною мовою, формується в результаті лінгвістичної обробки ПМТ.

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 відносин між ними з метою формування поняттєвої структури змісту вхідного тексту [1; 4].

Для відображення рольових відношень введено поняття неявного предикату. Під неявним предикатом в цьому разі розуміється відношення, яке не має відповідного лексичного еквіваленту в тексті. Так, наприклад, у словосполученні *фірма "Лінгвістика-93"* відношення "*мати назву*" між аргументами "фірма" та "Лінгвістика-93" відображається в тексті пробілом. З метою зберігання виразових засобів природно-мовного текстового представлення введені спеціальні засоби – префікси і постфікси предикатів і понять. Елементарна предикатна формула може містити також квантори єдності ($\forall$) та існування ($\exists$).

Елементарна предикатна формула має вигляд:

$$N\,P_k^q\,(\,LX_t^i,\,MY_g^j\,),$$

де *N, L, M* − відповідно префікси предиката та аргументів, які визначають тип семантичного класу; *P* − назва семантичного класу відношення; *X, Y* − назви семантичних класів понять. Аргументи мають фіксоване положення. Формула інтерпретується в термінах класичного числення предикатів: "поняття *X* знаходиться у відношенні *P* до поняття *Y*". Постфіксами виступають верхні та нижні індекси предикату і аргументів. Верхній індекс предикату q $(q\in Q)$ визначає лексико-граматичний спосіб мовного сполучення відношення і понять в тексті. Множина *Q* становить перелік як мовних одиниць (наприклад прийменники, частки тощо), так і граматичних ознак (наприклад відмінок управління між дієсловом і відповідним іменником), які відбивають правила сполучення відношень і понять в тексті. Нижній індекс предикату *k* визначає конкретний лексичний представник для відповідного семантичного класу *N*. Верхні індекси аргументів i та $j(i,j\in A)$ визначають граматичні характеристики понять (наприклад число, істота, неістота тощо). Множина *A* − це список граматичних характеристик понять, які виступають аргументами предикатної формули. Нижні індекси аргументів t і g $(t\in L,\ g\in M)$ визначають конкретний лексичний представник відповідних семантичних класів. У процесі формально-логічного виведення постфікси ігноруються. Вони є вирішальними на етапі синтезу опису фрагментів ПС природно-мовними засобами. В окремий лексико-семантичний клас відношень виділені лексичні одиниці, які мають значення модальності (*хотіти*, *вміти*, *треба*, *необхідно* тощо) [1; 5].

Уніфікація лексичних представників понять і відношень в рамках певного семантичного класу здійснюється за відношенням "*рід−вид*". Структура семантичного класу становить ієрархічну структуру, на верхньому рівні якої знаходяться найбільш загальні поняття (відношення), кожний нижчий рівень представлений поняттями (відношеннями), які конкретизують відповідні поняття (відносини) вищого рівня. Вибір родо-видового відношення для уніфікації понять і відношень в заданій ПГ має принципове значення. Оскільки заміна видових понять (відносин) на родові у вільних словосполученнях не веде до порушення семантичного значення вислову (чого не можна сказати про інші ієрархічні відношення).

Будь-яка ПГ визначається парадигматичними і синтагматичними відношеннями. Парадигматичні відносини ідентифікують системні зв'язки між поняттями в предметній галузі. Такі відносини, як правило, не відносять до конкретного тексту, оскільки там не реалізуються. Парадигматичні відношення фактично характеризують професійну компетентність фахівця.

Процес побудови моделі знань відбувається в декілька етапів. На першому етапі фахівець (експерт) укладає систему базових понять в заданій ПГ з відповідними прагматичними відносинами [4; 6]. Таку систему прийнято представляти у вигляді тезауруса. Призначення цього тезауруса − представити так звані "вертикальні" відносини між базовими поняттями, що існують в ПГ, і які не залежать від їх контекстного вживання.

Незалежно від ПГ можна виділити парадигматичні відносини, які характеризують систему, а не конкретну ПГ. Серед цих відносин виділяють такі: *частина, ціле*, *рід, вид*, *синонім*, *антонім, асоціативні відносини*. Асоціативні відносини

належать до слабко формалізованих відносин, можуть мати різний прагматичний зміст залежно від ПГ. Для кожного виду відносин в тезаурусі задаються окремі поля. Формат представлення тезауруса показаний в табл. 1.

Таблиця 1

Формат структури представлення 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 ПГ добре структурована, то код може містити закодовану семантичну інформацію про місце відповідного поняття в системі ПГ. Наприклад: 1.1.1 – означає що поняття знаходиться на третьому рівні ієрархії в системі. Якщо ПГ слабо формалізована, то доцільно ставити як код порядковий номер поняття в тезаурусі. Поле 2 містить саме поняття, яке представляється словом або словосполученням у початковій формі, щоб не плутати поняття з лексичною одиницею, 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 способом. Слід зазначити, що для конкретного поняття наповненість всіх полів не обов'язкова. Приклад заповнення наведений в таблиці 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 (ожеледиця на дорозі теж небезпека).

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

Синтагматичні відносини в ПГ визначають закономірності сполучуваності понять і відношень у певному тексті. Синтагматичну модель ПГ можна побудувати лише на підставі вивчення навчальної вибірки текстів заданої тематичної спрямованості, якщо йдеться 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 відповідної 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 ПГ за різномовними текстами полягає в тому, що відображення картини світу (ПГ) засобами мови в різних народів не збігається. Це пов'язане як із різними професійними поглядами на сутність явищ, фактів, способом доведення, так і об'єктивною різницею в самій картині світу (наприклад різні кліматичні умови тощо). Процес формування синтагматичної моделі продемонструємо на прикладі. Нехай за нашим запитом ми набрали декілька фрагментів різномовних текстів (рис. 1.). 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 непридатним для побудови синтагматичної моделі. Вимоги до насиченості тексту стосовно довжини/частоти визначає дослідник.

Українська мова

Національна безпека *- захищеність життєво важливих інтересів* ***людини*** *і* ***громадянина***, ***суспільства*** *і* ***держави***, *за якої забезпечуються 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 functioning 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"

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

Після побудови синтагматичної моделі корегується тезаурус, відбиває парадигматичні відносини в ПГ. Це пов'язане з тим, що класифікація знань про предметну галузь в різних мовах, як правило, не збігається. Так, *об'єктами* захисту в Україні є *людина*, *громадянин*, *суспільство*, *держава*, а в РФ – *особистість, суспільство*, *держава*.

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

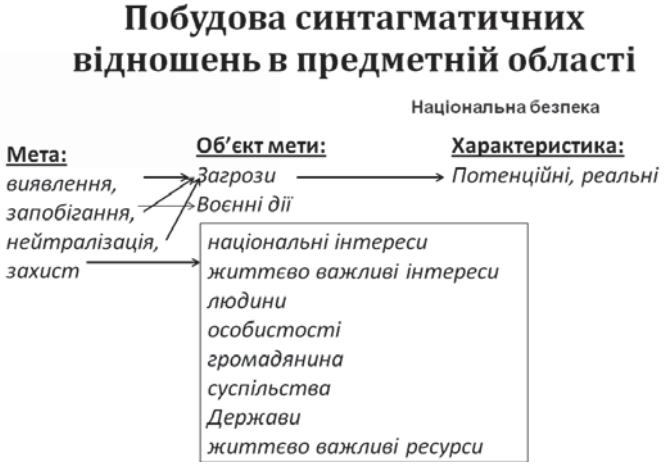

Рис. 2. Фрагмент синтагматичної моделі для поняття "національна безпека"

Формальна модель знань про предметну галузь на основі різномовних текстів заданої тематичної спрямованості призначена для усунення синтаксичної і семантичної омонімії в багатомовних системах машинного перекладу. Одвічною проблемою СМП є вирішення завдань багатозначності та визначення перекладних еквівалентів новими для системи слів. Запропонована модель дозволяє як розпізнавати нові слова, так і здійснювати вибір на множині можливих значень слів. Обробка нових слів відбувається таким чином. Якщо слово (лексичний відповідник поняття або відносини) відсутнє в перекладному словнику, то на основі накладання всіх його контекстів на предикатну структуру моделі можна обрати з відповідного лексико-семантичного класу представника з більш загальним значенням. Звичайно, щоб знайти відповідний еквівалент для відносин (поняття), треба мати для нового слова контекст, який визначається двома або більшою кількістю елементарних предикатних формул. Вже при визначених трьох елементарних предикатних формулах алгоритм працює з достовірністю до 70 %. Отже, при аналізі нових слів програма може порушувати стилістичну цілісність тексту, але зберігає його семантичну цілісність, що є вкрай важливим для розуміння тексту.

Запропонована модель представлення знань про предметну галузь дозволяє усувати синтаксичну омонімію при автоматичному перекладі, а також у випадку

появи в тексті нового поняття, відсутнього в словнику, автоматично добрати контекстний синонім до нового слова.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. *Замаруєва І.В.* Комп'ютерна модель розуміння природно-мовної текстової інформації / І.В. Замаруєва // Проблемы программирования. –1999. – № 2. С. 96–102.

2. *Иорданская Л.Н.* Синтаксическая омонимия в русском языке (с точки зрения автоматического анализа и синтеза) / Л.Н. Иорданская // НТИ. – 1967. – № 5. – С. 9–17.

3. *Колесников Н.П.* Омонимия в предложении и вопросы ее устранения / Н.П. Колесников. – М. : Наука, 1976. – 115 с.

4. Structure and maintenance of the linguistic processor in machine translation systems / Lenkov S.V., Zamarueva I.V., Balabin V.V., Pampuha I.V. // Вісник Черкаського державного техно-логічного університету. – 2009. – С. 141–143.

5. *Толубко В.В.* Задачі автоматичної обробки синтаксичної структури в знання-орієнтованій системі машинного перекладу / В.В. Толубко, О.О. Шипнівська, А.В. Ляшенко // Вісник Київського нац. ун-ту ім. Т. Шевченка. Серія : Військово-спеціальні науки. – 2010. – Вип. 27. – С. 136–140.

6. Технологічні аспекти реалізації автоматизованих систем машинного перекладу / В.В. Балабін, С.В. Лєнков, І.В. Замаруєва, І.В. Пампуха // Збірник наукових праць Військового інституту Київського національного університету імені Тараса Шевченка. – 2010. – № 26. – С. 55–64.

Отримано 12.03.2014