УДК 81’322.4 **Ніколаєвський О.Ю.** (КНУ)

ОСОБЛИВОСТІ АВТОМАТИЗОВАНОГО ФОРМУВАННЯ КОМПОНЕНТІВ ЗНАНЬ ПРО НАВКОЛИШНІЙ СВІТ В АРМ-ЕКСПЕРТ

У статті розглянуто лексичні засоби опису компонентів знань про навколишній світ. Наведено класифікаційну схему знань про навколишній світ (предметну галузь), які залучаються при машинному перекладі. Запропоновано засоби автоматизованого формування семантичної інформації до лексичних одиниць в частині так званих енциклопедичних знань, які в тексті не визначаються відповідними детермінантами.

Ключові слова: машинний переклад, лексичні засоби, знання про навколишній світ, лексичні детермінанти.

В статье рассмотрены лексические средства описания компонентов знаний об окружающем мире. Наведена классификационная схема знаний об окружающем мире (предметной области), которые необходимо привлекать при машинном переводе. Предложены средства автоматизированного формирования семантической информации, сопровождающей лексические единицы в части так называемых энциклопедических знаний, которые в тексте не сопровождаются соответствующими лексическими детерминантами.

Ключевые слова: машинный перевод, лексические средства, знания об окружающем мире, лексические детерминанты.

The article describes the lexical means of describing the components of knowledge about the world. Offered classification scheme of knowledge about the world (domain), that should be involved in machine translation. Proposed means for automatic forming semantic information, that has accompanying lexical units in terms of so-called encyclopedic knowledge, which in the text along with relevant lexical determinants.

Keywords: machine translation, lexical means, knowledge about the world, lexical determinants.

Вступна частина. На сьогодні побудова формальної моделі семантики тексту є самою слабкою ланкою в системах автоматичного опрацювання природно-мовної текстової інформації (ПМТІ). Однією з причин такого стану є різне бачення дослідників на обґрунтування і вибір одиниці змісту тексту. Вирішення цієї проблеми дозволить значно просунутися у галузі штучного інтелекту.

Пропонований підхід до розробки семантичної моделі ПМТІ, ґрунтується на таких концептуальних положеннях:

- вхідний природно-мовний текст – є зв'язний текст (тобто, дискурс);
- зв'язність дискурсу забезпечується графемними засобами оформлення тексту (наприклад, відношення взаємозв'язку між заголовками фрагментів тексту і змістом абзаців), лінгвістичними засобами (граматичними узгодженнями, анафоричними посиланнями тощо) і екстралінгвістичними (наприклад, часові, причинно-наслідкові зв'язки, які відповідають певній предметній області (ПО);
- всі ці засоби є інструментом кодування знань про світ (ПО);
- як елементи реального або абстрактного світу виступають його об'єкти, відображені в тексті у формі природно-мовних понять, відношень між ними та мовні характеристики понять і відношень;
- формування об'єктів за знаннями про світ може бути різним і залежить від цільової настанови прикладної задачі, проблемної області та логічної картини світу як носія інформації, так і того, хто її обробляє.

Формулювання цілей статті. Метою дослідження є обґрунтування й вибір семантичних одиниць тексту та визначення способу кодування семантичної інформації для одиниць, що представлені в різномовних текстів.

Вибір й обґрунтування семантичних одиниць тексту. Для вибору й обґрунтування семантичних одиниць ПМТІ проведено порівняльний аналіз відомих моделей семантики, які використовуються сьогодні в системах автоматичної обробки тексту. Порівняльну характеристику існуючих моделей розуміння ПМТ з точки зору рівнів організації тексту наведено в таблиці 1.

Таблиця 1

Порівняльний аналіз моделей семантики текстових об‘єктів

№ п/п	Моделі семантики ПМТ	Рівні подання тексту				
		Знаковий	Морфологічний	Синтаксичний	Семантичний	Прагматичний
1.	Модель "семантики переваги"	-	-	-	+	+ -
2.	Модель "концептуальних залежностей"	-	-	+	+	+ -
3.	Модель "смисл ↔ текст"	-	+	+	+	+ -
4.	Пропонована модель	графема	морфема	словосполучення	слово	поняття

Перша і друга моделі розроблялися для англійської мови, третя – для російської мови. Як видно з табл. 2.1. існуючі моделі розуміння ПМТ з точки зору розглядуваних семантичних одиниць тексту показують, що розпізнавання смислу в кращому випадку починається з морфологічного рівня мовної системи: модель "Смисл↔Текст" І. Мельчука [1]. В якості мінімальної одинці змісту в даній моделі пропонується морфема.

У відомих моделях семантики англійської мови, зокрема, в моделі "концептуальних залежностей" Р. Шенка [2] в якості мінімальної одинці змісту в даній моделі пропонується слово. В моделі "семантик переваги" Уілкса [3] аналіз тексту починається з рівня твердження, що в лінгвістиці відповідає простому ядерному реченню, фактично ігноруючи семантику морфемної структури англійської мови. Всі перераховані моделі ігнорують текст як знакову систему. В практичному плані аналіз знакового рівня організації ПМТІ обмежується відділенням синтаксичних розділових знаків від слова, виділенням абревіатур, скорочень тощо.

Аналіз текстів реальної складності показав, що вже на рівні знакової організації тексту людина використовує описові можливості семіотичної системи для кодування знань про фрагменти реальної дійсності. Так, використання лапок (наприклад, кінотеатр *"Салют"*) свідчить, що лексему в лапках не можна розглядати в значенні, поданому в словнику. Власні назви, наведені в тексті, можуть збігатися з написанням загально вживаних слів, але при цьому мати різний зміст (наприклад: депутат *Хмара*, прем‘єр-міністр *Major*, вул. *23 Серпня* тощо). Крім того, ряд лексем в тексті не підпорядковані граматичним правилам мови, а виступають як семантичні одиниці знакового рівня (наприклад: числа: *25,5 %, 10*, скорочення: *млн, кг*). Ці особливості ПМТІ й обумовили необхідність розробки знакового рівня організації тексту як початкового етапу побудови моделі розуміння тексту. Отже, з огляду на вище наведене, семантичний аналіз вхідного ПМТІ є розподіленим і здійснюється, починаючи з знакового рівня організації тексту. В якості мінімальної семантичної одиниці виступає ***графема*** (див. табл.1). Під графемою будемо розуміти “мінімальну смислову одиницю письмового тексту” [4, стор. 124].

На рівні організації тексту як лінгвістичної системи ми розглядаємо морфологічний рівень, синтаксичний та власне семантичний рівень мови.

На морфологічному рівні мови семантична модель представлена словотвірною моделлю. В якості мінімальної одиниці змісту виступає ***морфема*** (див. табл. 1). В розроблюваній системі машинного перекладу з усього арсеналу морфем розглядаються тільки префікси і суфікси.

На синтаксичному рівні мови семантична модель представлена стійкими словосполученнями і фразеологізмами. В якості мінімальної одиниці змісту виступає ***словосполучення*** (див. табл. 1). Основним завданням на цьому етапі є виділення термінів, понять, усталених словосполучень, які в тексті описані певною синтаксичною конструкцією і мають смислове навантаження, що не витікає із окремих слів синтаксичної конструкції.

На семантичному рівні мови основним завданням є переведення синтаксичної структури тексту до поняттєвої структури в термінах семантичних категорій. В якості мінімальної одиниці змісту виступає ***слово***, або нерозривне синтаксичне утворення, яке в розроблюваній системі машинного перекладу розглядається як одна одиниця перекладу (див. табл. 1).

На рівні організації тексту як системи відображення знань про фрагменти навколишнього світу семантична модель представлена моделлю знань про ПО. В якості мінімальної одиниці змісту виступає ***поняття***: в системі його парадигматичних і синтагматичних відношень (див. табл. 1).

Семантична модель на знаковому рівні організації тексту. Завдання автоматичного опрацювання ПМТІ на знаковому рівні організації тексту детально було розглянуто в [5]. В даній статті ми розглянемо лише склад інформаційного забезпечення семантичної моделі.

Як вже зазначалося, процедуру розпізнавання знань з ПО доцільно починати зі знакового (доморфемного) рівня організації тексту. Такий підхід зумовлений різноманітністю знакового (графемного) подання лексичних одиниць в тексті, яка визначає їх різні семантичні функції в тексті. Крім того, для вирішення задач перекладу суттєвим є також визначення структури тексту, для відокремлення службової інформації, виділення абзаців, заголовків тощо в тексті [6]. Текст при цьому розглядається як певним чином організована послідовність рядків і графем. Під графемою будемо розуміти мінімальну смислову одиницю письмового тексту. Задачею цього рівня розпізнавання є побудова формалізованого подання графемної структури тексту та розробка формального апарату виділення і класифікації текстових одиниць на множині рядків і графем.

Кінцевою метою доморфемного аналізу тексту є побудова графемної структури тексту, яка включає виділення на множині рядків і графем вхідного тексту таких семантично самостійних одиниць тексту: фрагментів, речень, синтагм, лексем; визначення класів перелічених одиниць тексту та встановлення відношень між ними в певному вхідному тексті.

Вхідними даними графемного аналізу є поточний текстовий файл і апріорні еталонні моделі (рядків і графем). В основу класифікатора графем покладені такі ознаки: тип знаку (цифра, буква, синтаксичний знак, службовий знак тощо), належність до алфавіту (латиниця, кирилиця, виключно російська, виключно українська), розмір (прописна, заголовна), фонетичні ознаки (голосна, приголосна).

На етапі інтерпретації доморфемної обробки тексту певні класи лексичних одиниць такі, як: ім'я, назва, позначення, абревіатура, скорочення тощо перевіряються на моделі знань про навколишній світ (ПО). Призначення цього етапу – виокремити класи лексем, які можуть збігатися за форматом представлення з класом мовних лексем. Так, наприклад, російське скорочення «*проц.*» за своїм написанням для системи може означати клас L1 – мовна лексема, а українська абревіатура «*омбр*» (окрема механізована бригада) за написанням збігається також з класом L1. Крім того, на цьому етапі для лексичних одиниць, для яких знайдено відповідники в моделі знань про ПО, записується і відповідна семантична інформація. Слід зазначити, що до моделі знань про ПО на доморфемному рівні, заносяться тільки ті одиниці, які або відсутні в перекладному словнику, або їх переклад призведе до семантичних помилок. Так, наприклад, українська назва парламенту – *Верховна Рада*, не

може перекладатися окремими словами на інші мови, теж відноситься і до прізвищ, імен тощо.

Інформаційне забезпечення семантичної моделі доморфемного аналізу включає словники, що відбивають екстралінгвістичні знання, необхідні для розпізнавання й вилучення знань про навколишній світ (ПО) безпосередньо із вхідного тексту або паперового словника. Формує такі знання фахівець в предметній галузі.

В основу єдиної семантичної параметризації словникових одиниць (для української, російської та англійської мов) закладені універсальні (енциклопедичні) знання про навколишній світ. Семантичний код – це двопозиційний цифровий код: перша позиція відбиває семантичний тип лексеми, друга – її семантичне значення.

Слід зазначити, що семантичний тип – це умовне розподілення лексичних одиниць на класи, що не перетинаються. Семантичний тип лексеми не залежить від графемного класу лексеми (класифікатор лексем представлений в попередньому звіті), так, скорочення і лексема з великої букви можуть входити до одного семантичного типу.

Так виділяються такі семантичні типи лексем:

1 – *географічна назва*. Цей клас включає просторові дані такі, як: назви міст, морів, океанів, річок, озер, материків тощо. Необхідність введення словника географічних назв обумовлена тим, що ці назви в тексті подаються без детермінуючих лексем, оскільки визначають енциклопедичні знання (тобто такі, що мають бути відомими). В цей клас ми не включили назви держав, оскільки для нашої ПО ці назви мають політичний контекст, це власне й обумовило внесення їх до іншого семантичного класу – політична назва;

2 – *історична назва*. Цей клас включає відомі назви історичних подій;

3 – *ім'я*. Необхідність введення даного семантичного типу обумовлена тим, що в англійських текстах мало відомі прізвища подаються разом із іменем. Це дозволяє, з одного боку, ідентифікувати, що це є особа й об'єднати дві лексеми в одне неподільне поняття, з іншого боку, визначення категорії роду для імені дозволяє досягти більшої точності при перекладі російською чи українською мовою;

4 – *установа*. Цей клас включає відомі назви організацій, установ, видів збройних сил тощо;

5 – *одиниця вимірювання*. Цей клас включає скорочення, що визначають одиниці вимірювання та назви місяців і днів тижня для англійської мови, які в тексті пишуться з прописної літери;

6 – *назва, що не перекладається*. Цей клас включає назви організацій, установ, прізвищ, назви вулиць тощо, які передаються засобами іншої мови виключно заданими правилами транслітерації;

7 – *посада*. Цей клас включає назви посад, які пишуться із заголовної літери;

8 – *політична назва*. Необхідність введення даного семантичного типу обумовлена тим, що назви держав в нашому контексті (ПО): воєнно-політичні тексти) розглядаються як геополітичні об'єкти, а не як географічні назви.

9 – *не визначений семантичний тип*. Даний семантичний тип призначається, коли лексема не підходить не під один із перерахованих семантичних типів. Якщо таких лексем набирається значна кількість, то класифікатор семантичних типів необхідно розширювати.

Друга позиція семантичного коду визначає семантичні характеристики лексеми у співставленні зі світом. Так виділяються такі значення семантичних типів лексем:

1 – *час*. Характеристика часу визначає лексему відповідного семантичного типу у часі;

2 – *простір*. Характеристика простору визначає лексему відповідного семантичного типу у просторі;

3 – *час-простір*. Дана характеристика притаманна деяким складним одиницям вимірювання (наприклад: км./год.);

4 – *кількість*. Характеристика, що відноситься виключно до оцінювання кількості;

5 – *об'єкт*. Характеристика, яка визначає конкретність (предметність) лексеми відповідного семантичного типу;

6 – *особа*. Характеристика яка визначає людину (посадову особу тощо);

7-8 – характеристики, які є резервними.

9 – *інше*. Характеристика лексеми відповідного семантичного типу, яка не підпадає під жоден із перерахованих класів.

В таблиці 2 наведено семантичну параметризацію словникових одиниць, які були виявлені при аналізі російських, англійських та українських текстів, спеціальної військової тематики.

Для автоматизації процесу формування словників екстралінгвістичних знань розроблено АРМ-«ЕКСПЕРТ». Розроблений програмний продукт підтримує англійську, російську та українську мову. Для кожної мови створюється окрема база даних, яка залучається у відповідності із мовою аналізованого тексту. Бази даних незалежно від мови мають єдину уніфіковану семантичну параметризацію (див. табл. 2), оскільки відбивають однакові фрагменти знань про навколишній світ. Формування словникових одиниць відбувається у 2-х режимах: безпосередньо за текстовим файлом та ручним введенням лексем оператором. За змістом текстові файли можуть загально довідкову інформацію (наприклад, словник імен, перелік одиниць вимірювання, словник географічних назв тощо), такі файли відбивають загально прийняті знання про навколишній світ і, як правило не супроводжуються пояснювальним контекстом. Наприклад: назви держав, відомих міст як правило не супроводжуються такими лексичними детермінантами як: (м.) *Москва* , (держава) *Україна,* (президент) *Clinton* тощо.

Таблиця 2

Семантична параметризація лексем на рівні організації тексту як знакової систем

Код семант. класу	**Приклади**	**Інтерпретація**
12!	*Asia, Africa, Europe, Средиземное море, Тихий океан*	Географічна назва: характеризується простором
21	*World War II*, *Перша світова війна*	Історична подія: характеризується часом
22!/21!	*Брестский мир*	Історична подія: характеризується часом і простором
35!	*Тарас, Martha, Александр*	Ім'я: особа
45!	*Рада національної безпеки і оборони, the National Security Council, the Central Intelligence Agency,*	Установа
51!	*January, Monday, хв.,р.*	Одиниця вимірювання: характеризується часом
52!	*См, кт, тт*	Одиниця вимірювання: характеризується простором
55!	*Омбр*	Одиниця вимірювання: структурний підрозділ
59!	*MHz*	Одиниця вимірювання: характеристика не визначена
65!	*Верховна рада, Дума,*	Власна назва, що транслітерується.
76!	*Президент, Верховний Головнокомандувач*	Посада: особа

Крім того, існують специфічні лексичні одиниці, які є загально прийнятими в заданій предметній галузі (наприклад: *омбр* – окрема механізована бригада). З цією метою в АРМ-«ЕКСПЕРТ» аналізується показова вибірка текстів заданої тематичної спрямованості і поповнюється база даних відповідної мови.

Слід зазначити, що у разі завантаження довільного природно-мовного тексту із заданої тематики до АРМ-«ЕКСПЕРТ» автоматично надходять лише слова з великої літери, абревіатури, скорочення, які за своїм написанням розпізнані на етапі доморфемного аналізу (наприклад: *млн., км./год., о-в, ртбр*), та слова підозрілі на скорочення чи інші класи, які не передаються на етап морфологічного аналізу. Автоматичне виявлення «підозрілих» слів в тексті досягається за рахунок того, що АРМ-«ЕКСПЕРТ» поєднаний з АРМ-«ПАРАДИГМА» і слова із тексту перевіряються спочатку перевіряються на базі всіх словоформ відповідної мови, що є в базі даних.

Висновки.

1. Модель семантики в розроблювальній системі машинного перекладу є розподіленою. Це пов'язане з тим, що текст розглядається нами як взаємодія трьох систем: знакової, мовної і системи знань про світ (ПО). Кожна з цих систем має свої (властиві лише їй) одиниці змісту і засоби формалізації семантики.

2. Мінімальною одиницею змісту (смислу) на знаковому рівні організації тексту є знак (графема), максимальною - рядок. В основу класифікатора знаків покладені такі ознаки: тип знаку (цифра, буква, синтаксичний знак, службовий знак тощо), належність до алфавіту (латиниця, кирилиця, виключно російська, виключно українська), розмір (прописна, заголовна), фонетичні ознаки (голосна, приголосна). Класифікатор рядків включає такі змістовно значущі класи: пустий рядок, повний рядок, неповний праворуч, неповний ліворуч, симетрично неповний.

3. Етап розпізнавання смислу на знаковому рівні дозволяє розв'язати наступні задачі: сформувати лексичні класи змістовно значущих понять в тексті; сформувати семантично правильні речення в тексті; сформувати змістовно закінчені фрагменти в тексті; визначити відношення між переліченими одиницями тексту, які проявляються на графемному рівні подання тексту.

4. Інформаційне забезпечення системи включає цілу низку словників, що відбивають екстралінгвістичні знання, необхідні для розпізнавання й вилучення знань з предметної області безпосередньо із вхідного тексту. Автоматизація формування компонентів інформаційного забезпечення семантичної моделі на знаковому рівні організації тексту забезпечує розроблений АРМ-«ЕКСПЕРТ».

ЛІТЕРАТУРА:

1. Мельчук И.А. Опыт теории лингвистических моделей «Смысл-Текст». Семантика, синтаксис / И.А. Мельчук – М.: Наука, 1974. – 314 с.

2. Шенк Р. Обработка концептуальной информации. / Р. Шенк – М.: Наука, 1980 – 360с.

3. Балабін В.В., Замаруєва І.В. Автоматизація когнітивного розпізнавання текстових об'єктів в умовах багатозначності і невизначеності // К.: ВІ КНУ. – Збірник наукових праць ВІ КНУ. – 2007. – №6, с.76-84.

4. Кондаков Н.И. Логический словарь-справочник / Н.И. Кондаков – М.: «Наука»., 1975. – 717 с.

5. Литвиненко Л.О. Особливості побудови лінгвістичного процесора доморфемного аналізу англійських військово-технічних текстів // Вісник Київського національного університету імені Тараса Шевченка. Військово-спеціальні науки. – К.: ВПЦ «Київський університет», 2012. – № 28. – С. 21-24.

6. Балабін В.В., Замаруєва І.В. Доморфемна обробка текстів в системах машинного перекладу // Збірник наукових праць Військового інституту Київського національного університету імені Тараса Шевченка, 2008. – № 11. – С. 78-84.

Рецензент: д.т.н., проф. Замаруєва І.В.