ВИКОРИСТАННЯ ОНТОЛОГІЇ ПРЕДМЕТНОЇ ОБЛАСТІ ДЛЯ УСУНЕННЯ НЕОДНОЗНАЧНОСТЕЙ ПРИ КОМП'ЮТЕРНОМУ ПЕРЕКЛАДІ ТЕХНІЧНИХ ТЕКСТІВ

Моренцова А.В.,
викладач кафедри англійської мови
технічного спрямування №1 факультету лінгвістики
Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського»

Комп'ютерний переклад (КП), який ще називають машинним перекладом (Machine Translation), є одним з напрямків комп'ютерної лінгвістики, яка досліджує використання програмного забезпечення для перекладу текстів з однієї природньої мови на іншу. Вперше ідею використання комп'ютерів для перекладу природніх мов запропонував англійський інженер-електрик Ендрю Бут в 1946 році.

На базовому рівні комп’ютерний переклад для пари мов виконувався шляхом послівного перекладу — простою заміною слів на одній природній мові словами іншої мови, що не давало якості перекладу тексту. Для більш-менш нормального перекладу необхідним був розгляд цілих фраз і всієї множини значень перекладаємого слова у вихідній мові. Одним з рішень цієї проблеми стало використання відмінностей у лінгвістичній типології, перекладу ідіом і ізоляції аномалій.

Існуючі системи КП допускають настроювання предметної області (ПрО), наприклад, таких як інформаційні технології, хімія, фізика, авіація й ін., до якої відноситься перекладаємий текст, поліпшуючи якість перекладу шляхом обмеження кількості припустимих замін перекладаємих слів. Цей спосіб особливо ефективний в областях, де використовується формальна або шаблонна мова. Із цього, наприклад, випливає, що комп'ютерний переклад державних і юридичних документів більш точний, ніж переклад розмови або менш стандартизованого тексту.

Ще одним способом, що поліпшують якість перекладу, є використання методів, заснованих на граматиці. Але для застосування цих методів потрібен кваліфікований лінгвіст, щоб ретельно проектувати граматику, яку ці методи використовують.

На сьогоднішній день у комп'ютерному перекладі однією з головних проблем є розв’язання ***лексичної багатозначності*** — неоднозначності змісту слова, коли у слова може бути більше ніж одне значення. Розв’язання лексичної багатозначності — це встановлення значення слова в деякому контексті [1]. І якщо для людини процес усунення багатозначності є підсвідомим і не представляє яких-небудь труднощів, то для комп'ютера розв’язання лексичної багатозначності являє собою досить складне завдання. У той же час вирішення цього завдання надто важливо для підвищення якості машинного перекладу [2].

Наявні підходи для подолання цієї проблеми можна розділити на дві групи: засновані на правилах й імовірнісні системи [3].

Системи, засновані на лінгвістичних правилах, виконують локальний або глобальний синтаксичний розбір. Такі системи передбачають всебічне знання слова і включають виконання великих граматичних і лексичних досліджень (аналізу тексту) для розв’язання багатозначності у вхідному тексті. Такий підхід ґрунтується на зовнішніх джерелах знань (knowledge-based methods), легко адаптується до будь-яких текстів і не прив'язаний до конкретної мови. При цьому слова перекладаються лінгвістичним способом – слово вхідною мовою заміняють найбільш підходящим словом вихідної мови з урахуванням ПрО. Засновані на правилах методи розбирають текст, створюючи проміжне, символьне представлення, від якого потім виробляється текст вихідною мовою. Але ці методи вимагають великих словників з морфологічною, синтаксичною й семантичною інформацією та великого зводу правил, що ускладнює їхню реалізацію.

Імовірнісні системи засновані на машинному навчанні й не використовують знання тексту. Вони просто застосовують статистичні методи до слів, що оточують неоднозначне слово. Алгоритми, засновані на даному підході показують досить гарні результати, однак вимагають навчання на

текстах, схожих з оброблюваними надалі, що пов'язане із проблемою розрідженості мови.

До останнього часу статистичні підходи були більш успішними, оскільки приблизно 90% середнього тексту відповідають цим простим умовам. Але є й інші 10%, що містять багатозначності, які потрібно розв’язати. Виходячи з досвіду виконаних авторами перекладів стандартів ISO/IEC із англійської на українську мову, у наукових і технічних текстах цей відсоток суттєво збільшується через те, що такі тексти містять значну кількість слів із множинним значенням — як специфічні терміни, так і загальної лексики.

Ідеальний підхід вимагає, щоб система комп'ютерного перекладу самостійно виконала все дослідження, необхідне для дозволу неоднозначності. Але існуючі й реалізовані на даний момент методи КП, на жаль, не дозволяють розв'язати це завдання. Одним з ефективних способів його розв'язку є застосування в комп'ютерному перекладі онтологій, зокрема онтології ПрО, до якої відноситься перекладаємий текст.

Онтологія – це формальне представлення знання, яке включає поняття (такі як об'єкти, процеси і т.д.) у ПрО й деякі відносини між ними. Але для цілей комп'ютерної лінгвістики більше підходить визначення, запропоноване Едвардом Хові : «Онтологія — це структура даних із заданими в ній символами, які дозволяють представляти концептуалізації для обробки комп'ютерними програмами» [4]. Якщо інформація, що зберігається, має лінгвістичну природу, можна говорити про словник. Онтології можуть використовуватися в якості джерела знання для систем КП. Основне їхнє призначення — забезпечити зв'язок між фрагментами тексту природньою мовою й поняттями ПрО (наприклад, класами або екземплярами онтології). З доступом до великої бази знань система може розв'язати багато двозначностей (особливо лексичних) самостійно.

Звичайно, виконуючи переклад, ми як люди можемо інтерпретувати ***фразу речення*** згідно з ***контекстом***, використовуючи наше знання, збережене в наших словниках. Система КП не в змозі диференціюватися між різними значеннями слова, оскільки синтаксис не

змінюється. З використанням досить великої онтології як джерела знання можуть бути зменшені можливі інтерпретації неоднозначних слів у певному контексті. Це пов'язане з тим, що контекст у конкретній ПрО тісно пов'язаний із семантикою (смисловим значенням) слів, а контекст і семантика найбільше повно відображаються саме онтологією цієї ПрО, представленою у вигляді онтологічної моделі. Семантичні описи є частиною онтології. При побудові онтологій максимально враховується семантика об'єктів, тому що семантичні описи відображають основні поняття об'єкта.

У нашім задачі онтологічна модель містить безліч понять (об'єктів) ПрО, які представляються словами вхідної мови. При цьому для багатозначних слів дається семантика кожного з можливих значень цього слова й ця семантика в комбінації зі словом представляє окреме поняття. У такий спосіб в онтологічній моделі можуть бути присутнім поняття позначені тим самим словом, але відмінні своєю семантикою. Другим елементом онтологічної моделі є відносини (зв'язки) між поняттями (словами). При цьому для кожної семантики багатозначного слова (по суті для кожного поняття) існує своє відношення (зв'язок) з деякою припустимою непустою множиною понять, кожне з яких так само визначене не тільки словом, але й семантикою. Відношення при цьому також має свою семантику, яка описує процес або дію.

Уведення в алгоритм комп'ютерного перекладу семантики дозволяє виконати семантичний аналіз структури речення на додаток до синтаксичного аналізу. Слід зазначити, що синтаксичний аналіз легко виконуються для мов, що піддаються типологічній класифікації на основі порядку слів у реченні (наприклад, англійська мова). Для флективних мов, для яких характерний відносно вільний порядок слів у реченні, завдання аналізу ускладнюється, але все-таки вирішується методом граматики залежностей [5]. При цьому розв'язок спрощується при використанні онтологічної моделі, що відображає залежності понять (слів речення). При семантичному аналізі виявляються основні значеннєві елементи речення, які в сукупності несуть зміст речення, що суттєво підвищує точність перекладу. Онтологія ПрО використовується для розпізнавання цих елементів як поіменованих сутностей. Семантична структура

речення отримується рекурсивним групуванням основних елементів речення за їхніми семантичними атрибутами. Онтологічна модель речення будується на основі семантичних значень елементів речення й відносин (структурним і семантичним) між ними.

Підводячи підсумок відзначимо, що використання в алгоритмах комп'ютерного перекладу онтології ПрО дозволяє усунути лексичну багатозначність і суттєво підвищити точність перекладу спеціалізованих технічних текстів.

Література:

1. *Agirre E., Edmonds P. G.* Word Sense Disambiguation: Algorithms and Applications. — U.K.:Springer, 2007. — 380 c.

2. *Турдаков Д.Ю.* Методы и программные средства разрешения лексической многозначности терминов на основе сетей документов : дисс. на соискание ученой степени канд. физ.-мат. наук: 05.13.11. / Московский государственный университет имени М.В. Ломоносова. — Москва, 2010. — 138 с.

3. *Сокирко А., Толдова С.* Сравнение эффективности двух методик снятия лексической и морфологической неоднозначности для русского языка. [Электронный ресурс]. Режим доступа: http://aot.ru/docs/RusCorporaHMM.htm

4. *Митрофанова О.А., Константинова Н.С.* Онтологии как системы хранения. [Электронный ресурс]. Режим доступа: http://window.edu.ru/resource/795/58795/files/68352e2-st08.pdf

5. *Касевич В. Б.* Структура предложения // Элементы общей лингвистики. — М.: Наука, 1977. — С. 91-92.