Перевод

Моренцова А. В.,
Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», м. Київ

ОНТОЛОГІЧНА МОДЕЛЬ ПРЕДМЕТНОЇ ОБЛАСТІ ЯК ІНСТРУМЕНТ ПОЛІПШЕННЯ ЯКОСТІ МАШИННОГО ПЕРЕКЛАДУ

Анотація: *Запропоновано новий підхід до розв'язання задачі усунення лексичної багатозначності, який базується на використанні розширеної онтологічної моделі предметної області при машинному перекладі технічних текстів. Цей підхід може використовуватися в алгоритмах машинного перекладу й суттєво підвищити точність і якість перекладу.*

Ключові слова: машинний переклад; онтологія; онтологічна модель; лексична багатозначність.

Ключевые слова: машинный перевод; онтология; онтологическая модель; лексическая многозначность.

Keywords: machine translation; ontology, ontological model; lexical ambiguity.

Машинний переклад (МП) — Machine Translation — є одним з напрямків комп'ютерної лінгвістики, яка досліджує використання програмного забезпечення для перекладу текстів з однієї природньої мови на іншу. Вперше використання комп'ютерів для перекладу природніх мов запропонував англійський інженер-електрик Ендрю Бут в 1946 році.

Перші моделі систем машинного перекладу були примітивними. 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. Граматика в традиційнім розумінні в них була відсутня повністю. Кінцева множина

лексичних одиниць із тексту, що перекладався, узгоджувалася з обмеженим контекстом з лексичних одиниць тексту перекладу [1]. Системи прямого перекладу не мали засобів вирішення проблем ***лексичної багатозначності***, не справлялися з незв'язаними мовними парами, не проводили ніякого лінгвістичного аналізу перед генеруванням перекладу, повторювали синтаксичні структури мови оригіналу, не ураховували мінімальних потреб синтаксичного й семантичного аналізу, і не встановлювали розрізнення частин мови, наприклад, іменників і дієслів [2].

Системи МП другого покоління, відрізнялися від систем першого покоління тим, що при їх проектуванні використовувалися модульні структури, що надавало можливість оновлення граматичних правил і словників, а також додавання нових мовних пар без негативного впливу на працездатність системи. Системи машинного перекладу третього покоління базувалися на використанні корпусної лінгвістики — сукупності текстів, зібраних відповідно до визначених принципів.

Наразі існуючі системи МП 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. Ще одним способом, що поліпшують якість перекладу, є використання методів, заснованих на граматиці.

Незважаючи на постійний розвиток і удосконалення методів та систем МП, на сьогоднішній день у машинному 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 [3]. 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.

Наявні підходи для подолання цієї проблеми можна розділити на дві групи: засновані на правилах й імовірнісні системи [4]. 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. При цьому слова перекладаються лінгвістичним способом – слово вхідною мовою заміняють найбільш підходящим словом вихідної мови з урахуванням ПрО.

Імовірнісні системи засновані на машинному навчанні й не використовують знання тексту. Вони просто застосовують статистичні методи до слів, що оточують неоднозначне слово. Імовірнісні системи, що використовують статистику спільної зустрічальності граматичних ознак слів у великих корпусах, потребують наявності загальнодоступного корпусу текстів, якого для української мови на даний час не існує. Створенням української онтологічної лексико-семантичної бази знань UkrWordNet (UWN) зараз займаються науковці Київського національного університету імені Тараса Шевченка [5].

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 автором перекладів стандартів ISO/IEC з 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.

Інші існуючі на даний час основні методи вирішення проблеми лексичної багатозначності представлені в [6], зокрема аналіз можливості використання для цього лексичної бази WordNet.

Але наразі всі наявні системи і методи МП не в змозі однозначно вирішити проблему неоднозначності, коли одному слову мови А може відповідати декілька слів мови В і навпаки. Одним з ефективних способів для розв'язання проблеми неоднозначності є застосування в машинному перекладі онтології ПрО, до якої відноситься текст, що підлягає перекладу.

Онтологія – це формальне представлення знання, яке включає поняття (такі як об'єкти, процеси і т.д.) у ПрО й деякі відношення між ними. Онтологію можна розглядати як базу знань спеціального виду з семантичною інформацією про визначену предметну область. Модель онтології ПрО містить визначені концепти (поняття, класи), властивості концептів (атибути, ролі), відношення між концептами (залежності, функції) і обмеження на використання, які визначаються аксіомами. Формальна модель онтології представляється у вигляді трійки множин O={T, R, F}, де T – множина понять ПрО; R – множина відносин між ними; F – множина функцій інтерпретації понять і відносин. Фундаментальні поняття визначеної ПрО відповідають класам онтології.

Модель, як правило, конкретизується в залежності від призначення і сфери застосування онтології. Для обробки інформації на природній мові, зокрема при машинному перекладі, необхідно застосування спеціалізованих онтологій. Основне їхнє призначення — забезпечити зв'язок між фрагментами тексту природньою мовою й поняттями ПрО [7]. Онтології можуть використовуватися в якості джерела знання для систем МП. З доступом до великої бази знань система може розв'язати багато двозначностей (особливо лексичних) самостійно.

Тому автор пропонує при застосуванні онтологій в машинному перекладі використовувати онтологічну модель ПрО, представлену в графічному вигляді (як приклад, ER-модель Чена) і дещо розширену під поставлену задачу.

Звичайно, виконуючи переклад, ми як люди можемо

інтерпретувати ***фразу речення*** згідно з ***контекстом***, використовуючи наше знання, збережене в наших словниках. Система МП не в змозі диференціюватися між різними значеннями слова, оскільки синтаксис не змінюється. З використанням досить великої онтології як джерела знань 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 і опосередковано — з відносинами між поняттями (словами), а контекст і семантика найбільш повно відображаються саме онтологією цієї ПрО, представленою у вигляді онтологічної моделі.

У нашому підході онтологічна модель містить множину 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 відношення (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. 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 і якість перекладу спеціалізованих технічних текстів.

Список використаної літератури:

1. *Whitelock P., Kilby K.* Linguistic and Computational Techniques in Machine Translation System Design, 2nd edn. – London: Ubiversuty College London Press, 1995.
2. *Hutchins W.J.* Linguistic Models in Machine Translation//UAE Papers in Linguistic. 1979. – 9. – P. 29-52.
3. *Agirre E., Edmonds P. G.* Word Sense Disambiguation: Algorithms and Applications. — U.K.:Springer, 2007. — 380 c.
4. *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
5. *Анісімов А.В., Марченко О.О., Ніконенко А. О.* UWN: Універсальна онтологічна база знань української мови // Проблеми програмування. — 2012. — №2-3. — С. 348-355.
6. *Романюк, А. Б., І. А. Сундутова, М. М. Романишин.* Методи вирішення лексичної багатозначності. Використання WORDNET для вирішення проблем багатозначності. // Вісник НУ «ЛП» «Комп'ютерні системи проектування. Теорія і практика» – 2011. – № 711. – С. 147–157.

7. *Лесько О.Н., Рогушина Ю.В.* Использование онтологии предметной области для снятия омонимии в естественно-языкових текстах // Проблеми програмування. — 2017. — №2. — С. 61-71.