УДК 81’255:004

Філологічні науки

Моренцова Алла Володимирівна
Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського»
(Київ, Україна)

ЗАСТОСУВАННЯ ОНТОЛОГІЧНИХ МОДЕЛЕЙ ПРЕДМЕТНОЇ ОБЛАСТІ ДЛЯ УСУНЕННЯ ЛЕКСИЧНОЇ БАГАТОЗНАЧНОСТІ ПРИ МАШИННОМУ ПЕРЕКЛАДІ

Анотація: *Розглянуті існуючі методи й способи усунення лексичної багатозначності, застосовувані в системах машинного перекладу. Запропонований новий підхід до розв'язання цієї задачі, який базується на використанні розширеної онтологічної моделі предметної області для усунення неоднозначностей при машинному перекладі технічних текстів. Цей підхід може використовуватися в алгоритмах машинного перекладу й суттєво підвищити точність перекладу.*

Ключові слова: *машинний переклад, лексична багатозначність, онтологія, онтологічна модель.*

Моренцова Алла Владимировна
Национальный технический университет Украины «Киевский политехнический институт имени Игоря Сикорского»
(Київ, Україна)

ПРИМЕНЕНИЕ ОНТОЛОГИЧЕСКИХ МОДЕЛЕЙ ПРЕДМЕТНОЙ ОБЛАСТИ ДЛЯ УСТРАНЕНИЯ ЛЕКСИЧЕСКОЙ МНОГОЗНАЧНОСТИ ПРИ МАШИННОМ ПЕРЕВОДЕ

Аннотация: *Рассмотрены существующие методы и способы устранения лексической многозначности, применяемые в системах машинного перевода. Предложен новый подход к решению этой задачи, который базируется на использовании расширенной онтологической модели предметной области для устранения неоднозначностей при машинном переводе технических текстов. Этот подход может использоваться в алгоритмах машинного перевода и существенно повысить точность перевода.*
Ключевые слова: *машинный перевод, лексическая многозначность, онтология, онтологическая модель*

Morentsova Alla
National Technical University of Ukraine
«Igor Sikorsky Kyiv Polytechnic Institute»
(Kyiv, Ukraine)

APPLICATION OF ONTOLOGICAL DOMAIN MODELS FOR ELIMINATION OF LEXICAL AMBIGUITY IN MACHINE TRANSLATION

Abstract: *The existing methods, applied in machine translation systems and ways of the solution to lexical ambiguity problem are considered. A new approach to the solution of this task which is based on using the expanded ontological domain model for elimination of ambiguity in machine translation of technical texts is offered. This approach can be used in algorithms of machine translation and can considerably increase the translation accuracy.*
Keywords: *machine translation, lexical ambiguity, ontology, ontological model.*

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. Ідея

автоматизації перекладу виникла майже одночесно з появою комп’ютерної технології..Вперше використання комп'ютерів для перекладу природніх мов запропонував англійський інженер-електрик Ендрю Бут в 1946 році. А вже у 1949 році, через п'ять років після запуску в експлуатацію в США першого досить потужного комп'ютера, відомий математик У. Уівер запропонував використовувати криптографічну техніку для механізації перекладу [1]. Незважаючи на неефективність технічних можливостей криптографії, ідея Уівера розбудила серйозний інтерес в багатьох лінгвістів і математиків до процесу механізації перекладу, і незабаром з'явилася потреба в особливій теорії, що вивчає проблематику машинного перекладу, а саме теорії машинного перекладу [2, 3].

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.

Слово в мові перекладу породжувалося від слова в мові оригіналу. Граматика в традиційнім розумінні в них була відсутня повністю. Речення конструювалися прямим заміщенням послідовності слів мови, з якої виконувався перклад, послідовністю слів мови, на яку виконувався переклад. Отже кінцева множина лексичних одиниць із тексту, що перкладався, узгоджувалася з обмеженим контекстом з лексичних одиниць тексту перекладу [4]. Однак загальне значення в різних мовах можуть виразити не тільки окремі слова, але й словосполучення. Тому, найпростіша система автоматичного перекладу повинна була шукати відповідності не тільки для окремих слів, але й для словосполучень, виконуючи так званий послівно-зворотній переклад. 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 [5].

У системах прямого перкладу процес перекладу розглядався лише як якась формальна заміна мовних одиниць тексту, що перекладався, на відповідні їм одиниці тексту перекладу, ігноруючи той факт, що суть перекладу полягає не в перекладі слів, а значень. Адже повідомлення – це ланцюжок сильно зв'язаних між собою значень, які у підсумку формують передану автором думку. А слова є лише покажчиками на семантичні поля даних значень з більш-менш семантичними ознаками і носіями граматичної інформації. Тому перші системи перекладу, розроблені за принципом «слово за словом», відсували синтаксичні та семантичні потреби на задній план і часто видавали на вихід переклад дуже низької якості. Це акцентувало складність мови й необхідність кращого аналізу й синтезу текстів.

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.

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. Корпус — це сукупність текстів, зібраних відповідно до визначених принципів. Ці принципи можуть бути як лінгвістичного, так і екстралінгвістичного характеру. Корпус може обмежуватися певними типами текстів, однієї або декількома різновидами якої-небудь мови, певним тимчасовим проміжком або комбінацією цих параметрів. Інформація про склад корпуса доступна дослідникові через інтерфейс корпуса (включаючи кількість слів у кожній категорії текстів і в корпусі в цілому,

інформацію про способи добору текстів, граматичні характеристики і т.д.). Об'єм паралельних корпусів ріс разом з дослідженнями в області обробки природньої мови.

При використанні корпусної лінгвістики застосовувалися дві методики машинного перекладу: статистичний підхід і метод перекладу за прикладами. У цих двох методиках лінгвістична інформація корпуса служить для породження нових перекладів. Усі системи перекладу по корпусній лінгвістиці використовують так звані опорні переклади, що містять тексти вхідною мовою і їх переклади. Еквівалентний переклад генерується за допомогою специфічного статистичного методу або добору прикладів, витягнутих з корпуса [6]. Обидва підходи не використовували лінгвістичні правила для аналізу текстів або вибору еквівалентів у мові, на яку виконується переклад.

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. 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.

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 [7].

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 [8].

Наявні підходи для подолання цієї проблеми можна розділити на дві групи: засновані на правилах й імовірнісні системи [9].

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. 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 [10].

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.

Ще одним підходом для правильного вибору еквіваленту слова, що перекладається стало застосування систем перекладу, що грунтуються на наступних двох принципах [11]:

1. На принципі вибору еквівалента по синтаксичній моделі вхідного тексту, найчастіше по синтаксичній моделі речення.

2. На принципі вибору еквівалента по семантичній моделі.

Обидві моделі звичайно застосовують у комплексі. Хоча такі системи машинного перекладу містять у собі й синтаксичний і семантичний аналіз, що робить їх більш надійними, ніж системи прямого перекладу, вони все ще допускають помилки. Причиною є те, що і самі моделі, і процедури вибору еквівалентів досить складні, і в них ще недостатньо ефективно проводиться граматичний аналіз для розв'язання синтаксичних і семантичних проблем.

Інші існуючі на даний час основні методи вирішення проблеми лексичної багатозначності представлені в [12], зокрема аналіз можливості використання для цього лексичної бази WordNet.

Але наразі всі системи і методи МП не в змозі однозначно вирішити проблему неоднозначності, коли одному слову мови А може відповідати декілька слів мови В і навпаки. У той же час ідеальний підхід вимагає, щоб система комп'ютерного перекладу самостійно виконала всі дослідження, необхідні для вирішення неоднозначності. 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 (наприклад, класами або екземплярами онтології) [13]. 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.

Як показує досвід автора, для значної кількості лінгвістів онтологія починається і закінчується використанням онтологічних словників. Дійсно, як зазначено в [14], сучасні онтологічні словники містять не лише конкретні значення слів, а й лексичні (антонімія, слова-відношення, номіналізація та ін.) та семантичні (гіперонімія/ гіпонімія, меронімія/голонімія та ін.) зв’язки між ними, що дає змогу використовувати їх для усунення багатозначності слів на основі цих зв’язків. Однак, онтологічні словники не описують в повній мірі мовні конструкції для вираження семантичних зв’язків між поняттями. 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. Така модель може бути більш або менш деталізованою, в деяких ПрО можуть виділятися ПрО нижчого рівня зі своїми онтологічними моделями, у яких звужується і деталізується коло понять у порівнянні з ПрО вищого рівня.

Наприклад, у ПрО «Хімія» можна виділити такі ПрО як «Хімічні технології», «Хімічне машинобудування», «Технології композитних матеріалів», «Неорганічна хімія» й под. Головне при машинному перекладі найбільш точно визначити належність тексту, що перекладається, до відповідної ПрО.

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 в 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 нашій задачі 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. 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 [15]. При цьому рішення 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.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ ТА ЛІТЕРАТУРИ:

1. Кво Ч. К. Технологии перевода. — М., 2008. – 256 с.
2. Tong L. C. translation: Machine-aided', in The Encyclopedia of Language and Linguistics, Vol. 9/R. E. Asher, J. M. Y. Simpson (eds). – Oxford: Pergamon Press, 1994. – P. 4730-4737.
3. Somers H. L. Machine Translation: History//Routledge Encyclopedia of Translation Studies/ M. Baker(ed.) – London: Routledge, 1998. – P. 140-143.
4. Whitelock P., Kilby K. Linguistic and Computational Techniques in Machine Translation System Design, 2nd edn. – London: Ubiversuty College London Press, 1995.

5. Hutchins W.J. Linguistic Models in Machine Translation//UAE Papers in Linguistic. 1979. – 9. P. 29-52.
6. Carl M. A Model of Competence for Corpus-based Machine Translation // in Proceedings of COLING 2000: Vol. 2. – Germany, 2000. [Електронний ресурс]. — Режим доступу: http://acl.upenn.edu/C/C00/C00-2145.pdf
7. *Agirre E., Edmonds P. G.* Word Sense Disambiguation: Algorithms and Applications. — U.K.:Springer, 2007. — 380 с.
8. *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.
9. *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
10. *Анісімов А.В., Марченко О.О., Ніконенко А. О.* UWN: Універсальна онтологічна база знань української мови // Проблеми програмування. — 2012. — №2-3. — С. 348-355.
11. Мирам Г.Э. Профессия: переводчик. 3-е изд. — М., 2004. — 160 с.
12. *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.
13. *Лесько О.Н., Рогушина Ю.В.* Использование онтологии предметной области для снятия омонимии в естественно-языкових текстах // Проблеми програмування. — 2017. — №2. — С. 61-71.
14. *Лозинська О.В., Давидов М.В.* Математична модель граматично-доповненої онтології // Вісник НТУ “ХПІ». — 2015. — №11(1120). — С.102-107

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