УДК 81'25:81'322.4
DOI https://doi.org/10.32782/bsps-2024.5.17

ОКРЕМІ АСПЕКТИ ВИКОНАННЯ МАШИННОГО ПЕРЕКЛАДУ НА ОСНОВІ ПРАВИЛ

Катерина Рябова
кандидат юридичних наук,
старший викладач кафедри англійської мови
Державного університету інформаційно-комунікаційних технологій
вул. Солом'янська, 7, м. Київ, Україна
orcid.org/0009-0006-6455-3845
e-mail: rjabova3107@gmail.com

Анотація. *У статті проведено загальний аналіз машинного перекладу на основі правил, виокремлено певні правила, якими керуються дослідники й розробники під час розроблення сучасних систем машинного перекладу. Розглянуто історичний досвід архітектури МП, що включає перше покоління (1960–1980-і роки) й базувалося на прямому перекладі, друге покоління (1980-і роки до теперішнього часу) складається із систем на основі правил, таких як системи передачі й інтерлінгва, а третє покоління (1990-і роки до теперішнього часу) включає системи на основі корпусів, які базуються або на статистичних даних, або на основі прикладів. Обґрунтовано, що машинний переклад на основі правил (Rule-based Machine Translation), також відомий як машинний переклад на основі знань (Knowledgebased Machine Translation), спирається на морфологічні, синтаксичні, семантичні й контекстуальні знання про вихідну та цільову мови відповідно і зв'язки між ними для виконання завдань перекладу. Доведено, що в машинному перекладі на основі правил лінгвіст формалізує лінгвістичні знання в лексиконах і правилах граматики. Ці знання використовуються системою для аналізу речень мовою оригіналу та їх перекладу. Розглянуто окремі системи МП на основі правил, такі як Systran, Lucy LT, Apertium, ParSit, і деякі системи, які концентруються виключно на перекладі окремих діалектів. Аналіз цих систем виявив, що, якщо вихідна мова багата морфологічно, для аналізу вихідного тексту використовуються специфічні мовні морфологічні правила. Потім застосовуються специфічні для мови синтаксичні правила для визначення синтаксичних категорій слів, що містяться в реченні. Зроблено висновок, що машинний переклад на основі правил – це парадигма машинного перекладу, у якій лінгвістичні знання кодуються експертом у формі правил, що перекладаються з вихідної мови цільовою.*

Ключові слова: *машинний переклад, автоматизований переклад, теорія машинного перекладу, переклад на основі правил, синтаксичні правила, синтаксичні категорії слів.*

SOME ASPECTS OF RULE-BASED MACHINE TRANSLATION

Kateryna Riabova
PhD. in Law,
Senior Lecturer at the Department of English Language
State University of Information and Communication Technologies
Solomianska Str., 7, Kyiv, Ukraine
orcid.org/0009-0006-6455-3845
e-mail: rjabova3107@gmail.com

Abstract. *The article provides a general analysis of rule-based machine translation and identifies certain rules that guide researchers and developers in the development of modern machine translation systems. The historical experience of MT architecture is considered, which includes the first generation (1960s–1980s) based on direct translation, the second generation (1980s to the present) consisting of rule-based systems such as transfer and interlingual systems, and the third generation (1990s to the present) including corpus-based systems based on either statistical data or examples. It is substantiated that Rule-based Machine Translation, also known as Knowledgebased Machine Translation, relies on morphological, syntactic, semantic and contextual knowledge of the source and target languages, respectively, and the relations between them to perform translation tasks. It is proved that in rule-based machine translation, a linguist*

formalises linguistic knowledge in lexicons and grammar rules. This knowledge is used by the system to analyse sentences in the source language and translate them. Some rule-based MT systems, such as Systran, Lucy LT, Apertium, ParSit, and some systems that focus exclusively on the translation of certain dialects, are considered. Analysis of these systems has shown that if the source language is morphologically rich, language-specific morphological rules are used to analyse the source text. Then, language-specific syntactic rules are applied to determine the syntactic categories of words contained in the sentence. It is concluded that rule-based machine translation is a machine translation paradigm in which linguistic knowledge is encoded by an expert in the form of rules that are translated from the source language into the target language.

Key words: *machine translation, automated translation, machine translation theory, rule-based translation, syntactic rules, syntactic word categories.*

З часом розвиток машинного перекладу (далі – МП) призвів до появи кількох типів систем машинного перекладу, кожен із яких має свої сильні та слабкі сторони. МП на основі правил є одним із найпоширеніших типів МП поряд зі статистичним і нейронним. Метою статті є проведення загального аналізу МП на основі правил, а також виокремлення певних правил, якими керуються дослідники й розробники використовуючи «правила» під час розроблення сучасних систем МП окремих термінів, таких, наприклад, як автоматичний і машинний переклад тощо. В основному вчені-лінгвісти вивчають проблематику МП як цілісної парадигми, рідко приділяючи глибоку увагу окремим типам МП.

Незважаючи на те що дослідженню проблем МП на основі правил приділена недостатня увага вчених, варто виділити таких дослідників, які займалися RBMT останні два десятиліття, і їхня увага засвідчена численною кількістю лінгвістичних праць. Так, зокрема, варто виділити праці таких учених: С. Срилекха [2], Д. Торрегроса [1], Н. Пасрича [1], Б. Р. Чакравартi [1], Т. А. Пірінен [3], Б. Баніц [4], Х. Сомерс [5], П. Чароенпорнсават [6], В. Сорнлертламванич [6], Т. Чароенпорн [6], М. А. Сгайера [7], М. Зригі [7] й ін.

Окремо варто проаналізувати в загальному вигляді дослідження й результати таких науковців. Д. Торрегроса, Н. Пасрича, Б. Р. Чакравартi, М. Масуд та М. Аркан провели масштабне дослідження, де описували різні підходи до використання інформації, що міститься в системах МП на основі правил (Rule-based Machine Translation, скорочено RBMT), для покращення системи на основі корпусу, а саме моделі нейронного МП, з акцентом на сценарій із низьким ресурсом [1, с. 126].

С. Срилекха наголошує на поєднанні ідей і методів з лінгвістики, інформатики, штучного інтелекту, теорії перекладу і статистики для автоматизації процесу перекладу з однієї мови іншою. На його думку, основними труднощами в МП є різниця між вихідною та цільовою мовами та їх неоднозначність. Проведена ним класифікація підходів RBMT, МП на основі перекладу, МП на основі інтерлінгви й МП на основі словника надалі буде широко використовуватися під час досліджень RBMT [2].

Т.А. Пірінен, аналізуючи наявні дані FinnWordNet, Omorfi й Apertium-eng, створив правила лексичного вибору та перекладу вручну, розробив словники й правила на основі спільних даних завдань, описав використання спільних даних завдань як своєрідний робочий процес розробки, керований тестуванням, у розробці RBMT, показав, що він ідеально підходить для сучасного робочого процесу безперервної інтеграції RBMT у розробці програмного забезпечення й забезпечує значне підвищення балів BLEU з мінімальними зусиллями [3, с. 338].

Д. Торрегроса, Н. Пасрича, Б. Р. Чакравартi, М. Масуд та М. Аркан провели експеримент на основі системи RBMT Lucy LT з таких мовних пар: англо-іспанської (загальний і медичний домен), англійської баскської, англійської ірландської та англійської спрощеної китайської мови в сценарії з недостатнім ресурсом, використовуючи корпуси з приблизно мільйоном паралельних записів. Результати свідчать про те, що додавання морфологічної інформації до мови оригіналу є таким же ефективним, як і використання підслівних одиниць у цій конкретній ситуації [1, с. 128].

Існує три покоління архітектури МП: перше покоління (1960–1980-і роки) базувалося на прямому перекладі, друге покоління (1980-і роки до теперішнього часу) складається із систем на основі правил, таких як системи передачі й інтерлінгва, а третє поко-

ління (1990-і роки до теперішнього часу) включає системи на основі корпусів, які базуються або на статистичних даних, або на основі прикладів. Водночас системи прямого перекладу використовували дослівний переклад без чіткого вбудованого лінгвістичного компонента, системи, основані на правилах і на основі корпусів, є набагато складнішими [4, с. 55].

МП на основі правил (Rule-based Machine Translation), скорочено RBMT, також відомий як машинний переклад на основі знань (Knowledgebased Machine Translation), спирається на морфологічні, синтаксичні, семантичні й контекстуальні знання про вихідну та цільову мови відповідно і зв'язки між ними для виконання завдань перекладу. Лінгвістичні знання допомагають системам МП здійснювати переклад через доступні комп'ютерам словники та правила граматики, основані на теоретичних лінгвістичних дослідженнях. Цей раціоналістичний підхід контрастує з емпіричним підходом, який розглядає процес перекладу як імовірнісну подію й тому містить статистичні моделі перекладу, отримані з мовних корпусів.

Іншими словами, переклад на основі правил передбачає застосування морфологічних, синтаксичних і/або семантичних правил до аналізу тексту вихідної мови й синтезу тексту цільової мови, що вимагає лінгвістичних знань як вихідної, так цільової мови, а також розуміння відмінностей між ними [4, с. 58].

У машинному перекладі на основі правил лінгвіст формалізує лінгвістичні знання в лексиконах і правилах граматики. Ці знання використовуються системою для аналізу речень мовою оригіналу та їх перекладу. Хоча цей підхід не вимагає жодних навчальних корпусів і надає контроль над перекладами, створеними системою, процес кодування лінгвістичних знань потребує багато часу експерта. Яскравими прикладами систем RBMT є оригінальна, основана на правилах система Systran, Lucy LT та Apertium. Натомість системи машинного перекладу на основі корпусів навчаються перекладати з прикладів, як правило, у формі вирівняних корпусів на рівні речень. З одного боку, цей підхід, як правило, більш дорогий з погляду обчислень і пропонує обмежений контроль над створеними перекладами. Крім того, це неможливо для мовних пар, які мало мають або взагалі не мають доступних паралельних ресурсів. З іншого боку, він може похвалитися набагато вищим охопленням цільової мовної пари залежно від наявності паралельних корпусів. Прикладами парадигм МП на основі корпусу є статистичний переклад на основі фраз і моделі нейронного машинного перекладу (NMT) [1, с. 130].

Підхід, оснований на правилах, є першою стратегією, якої дотримувалися дослідники МП. Оскільки правила пишуться людиною на основі її лінгвістичних знань, очевидно, що така людина може глибоко аналізувати текст як на синтаксичному, так і на семантичному рівнях. Однак тут прослідковуються дві проблеми: по-перше, необхідною є наявність великих лінгвістичних знань, по-друге, неможливо написати правила, які будуть охоплювати всю мову.

Системи МП, основані на правилах, далі поділяються на трансферальні й інтерлінгвальні (Interlingua) підсистеми. Системи Interlingua працюють з абстрактним проміжним представленням вихідного тексту, з якого генерується цільовий текст [4, с. 68].

Системи передачі (трансферальні) є більш поширеним підходом до МП на основі правил. Системи, основані на передачі, аналізують речення вихідного тексту за реченням, ідентифікуючи частину мови кожного слова та його можливі значення [5, с. 431].

Якщо вихідна мова багата морфологічно, для аналізу вихідного тексту використовуються специфічні мовні морфологічні правила. Потім застосовуються специфічні для мови синтаксичні правила з метою визначення синтаксичних категорій слів, що містяться в реченні. Нарешті, система визначає цільове слово й генерує цільове речення, точно дотримуючись структури вихідного речення, далі піддаючи цільове речення простій процедурі морфологічного генерування, щоб застосувати специфічні для цільової мови морфологічні правила до виводу МП [4, с. 61].

Щоб покращити якість МП на основі правил, потрібно змінювати або додавати деякі правила генерації чи аналізу. Цей метод вимагає великих лінгвістичних знань, і не можна гарантувати, що точність буде кращою. Наприклад, у разі модифікації деяких правил це не лише змінює неправильні речення на правильні речення, а й може негативно впливати на правильні речення.

Аналізуючи вищевикладене, можна дійти висновку, що використання знань RBMT є корисним для покращення продуктивності систем NMT у сценарії з недостатнім ресурсом.

Кожен МП на основі правил має власну стратегію перекладу. Наприклад, П. Чароенпорнсават, В. Сорнлертламванич і Т. Чароенпорн у 2002 році, коли МП на основі правил був більш поширеним, оскільки нейронний МП тільки починав формуватися, аналізуючи як практичний приклад систему ParSit (ParSit – це машинний переклад з англійської мови тайською з використанням міжмовного підходу), зазначають, що ParSit складається з чотирьох модулів: модуля аналізу синтаксису, модуля семантичного аналізу, модуля генерації семантики та модуля генерації синтаксису [6]. Разом із тим під час виконання перекладу системою ParSit виділялося дві основні групи помилок: перша група – неправильне значення (відсутні деякі слова; наявні «зайві» слова; використання неправильного слова), друга – неправильне впорядкування.

У системі RBMT Lucy LT лінгвістичні знання формалізовані лінгвістами у вигляді комп'ютерних граматик, одномовних і двомовних лексиконів. Одномовні лексикони – це збірки лексичних статей; кожна лексична стаття є набором пар ознака-значення, що містить морфологічну, синтаксичну й семантичну інформацію. Записи двомовної лексики включають лексичні відповідності джерело-ціль і за бажанням контекстуальні умови та дії. Граматики – це колекції перетворень до анотованих дерев. Система Lucy LT поділяє процес перекладу на три послідовні фази: аналіз, передача й генерація. Під час фази аналізу вихідне речення токенізується та морфологічно аналізується за допомогою лексикону, який ідентифікує кожну поверхневу форму та всі її правдоподібні морфологічні значення. Далі аналізатор діаграм Lucy LT разом із граматикою аналізу, що складається з доповнених синтаксичних правил, витягує основну структуру дерева синтаксису й коментує її. Граматики передачі та генерації потім послідовно застосовуються до цього дерева, яке піддається численним анотаціям і перетворенням, що додають інформацію про еквівалентності в цільовій мові й адаптують оригінальні структури вихідної мови до відповідних у цільовій мові. Нарешті, кінцеві вузли дерева генерації збираються в перекладене речення [1, с. 130].

Окремий інтерес становлять системи, що обробляють діалекти. Ця потреба випливає головним чином із величезного діалектного вмісту, доступного в Інтернеті, який можна використовувати в різних сферах, у тому числі в різноманітних галузях досліджень поведінки користувачів, таких як системи рекомендацій, веб-майнінг або аналіз настроїв тощо. Більшість із цих діалектів є в Інтернеті через дописи в блогах, обговорення на форумах або взаємодії користувачів у соціальних мережах, таких як Facebook і YouTube.

Наприклад, існує система RBMT Elissa, яка різною мірою підтримує переклад арабських діалектів, таких як єгипетський, іракський, левантійський і мови Перської затоки, арабською мовою. Система ідентифікує вихідний діалект із подальшим морфологічним аналізом, виконаним за допомогою аналізатора MADA. Потім вона генерує остаточний текст цільовою мовою на основі діалектів, унесених до словників MSA (Modern Standard Arabic) і попередньо реалізованих правил морфологічного перенесення [7, с. 312].

Система RBMT потребує великих людських зусиль для підготовки правил і лінгвістичних ресурсів, таких як морфологічні аналізатори, тегери частин мови й синтаксичні аналізатори, двомовні словники, правила перенесення, морфологічний генератор, правила зміни порядку тощо. Системи МП на основі правил працюють на базі специфікації правил морфології, синтаксису, лексичного відбору та перенесення. Збірник правил і двомовний або багатомовний лексикон є ресурсами, які використовуються в RBMT [2].

Модель передачі в RBMT включає три етапи: аналіз, передачу й генерацію. Під час фази аналізу виконується лінгвістичний аналіз вхідного-вихідного речення, щоб отримати інформацію щодо морфології, частин мови, фраз, контекстуальної сутності слів та усунення їх неоднозначності. На етапі лексичного перенесення є два кроки, а саме: переклад слів і переклад граматики. У перекладі слів кореневе слово вихідної мови замінюється кореневим словом цільової мови за допомогою двомовного словника, а в граматичному перекладі перекладаються суфікси.

На етапі генерації відбувається узгодженість роду, числа й особи локальних груп фраз, а також рід дієслів [2].

Окремою проблемою RBMT є велика кількість низькочастотних синонімів у перекладах під час фази лінгвістичного аналізу. Цікаве рішення запропонував Т.А. Пірінен на прикладі фінської мови. Для аналізу він узяв словник Wordnet і на основі експерименту спробував увести автоматичні механізми створення правил для уточнення перетвореного словника. Для автоматичного завантаження правил лексичного вибору використано дані Europarl corpus та оновлено лексичний вибір деякими ручними правилами, які або не охоплюються зверненнями Europarl, або спотворені неправильно. Таким чином, фазу «навчання» в розробці RBMT замінено дуже простим напівавтоматичним робочим процесом проєкту, виконуваним носієм мови, який складається з такого: 1) зібрати всі лексеми, невідомі словнику вихідної мови, і додати до них необхідну морфолексичну інформацію. 2) зібрати всі лексеми, невідомі словнику двомовного перекладу, і додати їхні переклади. 3) зібрати всі лексеми, невідомі словнику цільової мови, і додати їх до словника з необхідною морфолексичною інформацією [3, с. 339].

Можемо зробити висновок, що розроблена Т.А. Пірінен напівавтоматизація полягає в збиранні різних невідомих лексем або одиниць поза словниковим запасом, а також спробі вгадати лексичний запис або кілька правдоподібних записів для них, а автор словника чи перекладач вибере й виправить їх.

Отже, машинний переклад на основі правил – це парадигма машинного перекладу, у якій лінгвістичні знання кодуються експертом у формі правил, що перекладаються з вихідної мови цільовою. Хоча цей підхід забезпечує повний контроль над вихідним текстом, вартість формалізації необхідних лінгвістичних знань набагато вища, ніж навчання системи на основі корпусу, де використовується підхід машинного навчання для автоматичного навчання перекладу з прикладів.

ЛІТЕРАТУРА

1. Torregrosa D. та ін. Leveraging rule-based machine translation knowledge for under-resourced neural machine translation models. *Proceedings of MT Summit XVII.* Dublin, 2019. № 2. С. 125–133. URL: https://aclanthology.org/W19-6725.pdf.
2. Sreelekha S. Statistical vs rule based machine translation; a case study on Indian language perspective. URL: https://arxiv.org/pdf/1708.04559.
3. Pirinen T.A. Apertium-fin-eng – rule-based shallow machine translation for WMT 2019 shared task. *Proceedings of the Fourth Conference on Machine Translation (WMT).* Florence, 2019. August 1–2. С. 335–341. URL: https://statmt.org/wmt19/pdf/53/WMT36.pdf.
4. Banitz B. Machine translation: a critical look at the performance of rule-based and statistical machine translation. *Cad. Trad., Florianópolis.* January, 2020. С. 54–71. URL: https://doi.org/10.5007/2175-7968.2020v40n1p54.
5. Somers H. Machine translation: latest developments. The Oxford handbook of translation studies / Malmkjaer, Kirsten ; Windle, Kevin (Ed). Oxford : Oxford University Press, 2011. С. 427–440. URL: https://personalpages.manchester.ac.uk/staff/harold.somers/Mitkov-book-chapter.pdf.
6. Improving translation quality of rule-based machine translation / P. Charoenpornsawat et al. *COLING-02: Machine Translation in Asia.* 2002. URL: https://aclanthology.org/W02-1605.pdf.
7. Sghaiera M.A., Zrigui M. Rule-Based Machine Translation from Tunisian Dialect to Modern Standard Arabic. 24th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems. Mohamed Ali Sghaier et al. Procedia Computer Science, 176 (2020). С. 310–319. URL: https://www.sciencedirect.com/science/article/pii/S1877050920318573.

REFERENCES

1. Torregrosa, D. et. (2019). Leveraging rule-based machine translation knowledge for under-resourced neural machine translation models. Proceedings of MT Summit XVII. № 2. Dublin, Aug. 19-23, Dublin, 2019. С. 125–133. https://aclanthology.org/W19-6725.pdf.
2. Sreelekha, S. Statistical vs rule based machine translation; a case study on Indian language perspective. https://arxiv.org/pdf/1708.04559.
3. Pirinen, T. A. (2019). Apertium-fin-eng – rule-based shallow machine translation for WMT 2019 shared task. Proceedings of the Fourth Conference on Machine Translation (WMT), Florence, August 1-2, 2019. Florence, С. 335–341. https://statmt.org/wmt19/pdf/53/WMT36.pdf.

4. Banitz, B. (2020). Machine translation: a critical look at the performance of rule-based and statistical machine translation. Cad. Trad., Florianópolis, January, 2020. Florianópolis, C. 54–71. https://doi.org/10.5007/2175-7968.2020v40n1p54.
5. Somers, H. (2011). Machine translation: latest developments. The Oxford handbook of translation studies / Malmkjaer, Kirsten; Windle, Kevin (Ed). Oxford: Oxford University Press, C. 427–440. https://personalpages.manchester.ac.uk/staff/harold.somers/Mitkov-book-chapter.pdf.
6. Charoenpornsawat, P. et. (2002). Improving translation quality of rule-based machine translation. COLING-02: Machine Translation in Asia. https://aclanthology.org/W02-1605.pdf.
7. Sghaiera, M.A., Zrigui, M. (2020). Rule-Based Machine Translation from Tunisian Dialect to Modern Standard Arabic. 24th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems. Mohamed Ali Sghaier et al. / Procedia Computer Science 176 C. 310–319. https://www.sciencedirect.com/science/article/pii/S1877050920318573.