УДК81’32.2.25

Фокін С.Б., к. філол. н., доц.
Інститут філології КНУ імені Тараса Шевченка, м. Київ

ЗАСОБИ АВТОМАТИЗОВАНОГО ЗБАГАЧЕННЯ КОМП’ЮТЕРНОЇ ПЕРЕКЛАДАЦЬКОЇ ПАМ’ЯТІ

У статті розглядається алгоритм розробленої автором програми, що дозволяє збагачувати комп’ютерну перекладацьку пам’ять і, відтак, поліпшувати якість машинного перекладу; збагачення перекладацької пам’яті передбачає збільшення кількості потенційних еквівалентів речень, словосполучень та нетермінологічих лексичних одиниць. У той час як більшість сучасних програм спрямовані на укладання перекладацької пам’яті з цілих речень, а також термінологічної бази даних, при цьому проміжна ланка (частини складних речень, нетермінологічні сталі словосполучення) в роботі програми машинного перекладу залишаються незадіяними.

Ключові слова: *автоматична вибірка, терміни, комп’ютерний переклад, алгоритм, закономірні відповідники.*

Дедалі більшого поширення набувають програми, що автоматично опрацьовують текстові масиви, укладають глосарії, генерують словники, здійснюють вибірку термінів. Це й різноманітні конкорданси, лексикографічні онлайн-ресурси тощо. Приміром, «Електронний словник мови Тараса Шевченка» [1], «Словник порівнянь Юрія Андруховича та Оксани Забужко» [2] та інші. Переважно автоматичними на сьогодні є частотні словники, до прикладу, Іспанська королівська академія мови надає можливість на сайті згенерувати список з 1000, 5000, 10000 найчастотніших словоформ [3]. У свою чергу програми перекладу дедалі поліпшують свою якість, хоча й залишають бажати кращого. Розглянемо два приклади:

(1) Todos los Miembros de las Naciones Unidas son ipso facto partes en el Estatuto de la Corte Internacional de Justicia. Un Estado que no sea Miembro de las Naciones Unidas podrá llegar a ser parte en el Estatuto de la Corte Internacional de Justicia, de acuerdo con las condiciones que determine en cada caso la Asamblea General a recomendación del Consejo de Seguridad. [4]

Всі члени Організації Об’єднаних Націй є в силу самого факту учасниками Статуту Міжнародного Суду. Держава, яка не є членом Організації Об’єднаних Націй, може стати учасницею Статуту Міжнародного Суду, відповідно до умов, визначених у кожному випадку Генеральна Асамблея за рекомендацією Ради Безпеки [Переклад Google Translate].

(2) Y en prueba de conformidad y aceptación, firman el presente document por duplicado ejemplar y a un solo efecto, en el lugar y fecha al principio indicados [з договору].

І в тестуванні відповідності та приймання, підписали цей документ у двох примірниках і один ефект, в тому місці і зазначеної дати [Переклад Google Translate].

Як видно з прикладів, широко відомі документи, які перекладені багатьма мовами світу, очевидно, вже потрапили до перекладацької пам’яті. Перший фрагмент, за винятком одного порушення відмінка, перекладено майже бездоганно. Натомість

другий приклад у перекладі виявляється не лише неузгодженим, граматично некоректним, а й навіть приблизно незрозумілим. Отже, резерви для вдосконалення машинного перекладу ще безсумнівно наявні.

Переважна більшість сучасних програм машинного перекладу працює на основі використання так званої перекладацької пам'яті – масиву оригіналів і перекладів, звідки комп'ютер підбирає еквіваленти. Якщо речення оригіналу збігається з наявним у пам'яті більш, ніж на 70%, користувачу програми пропонується переклад, що зберігається в перекладацькій пам'яті. Втім, 70% збігу не означає збіг 70% слів. Скажімо, одне переставлене на іншу позицію слово може знизити ступінь збігу одразу до 70%. Певна річ, половина речення або ж словосполучення, різного роду звороти (дієприкметникові, інфінітивні), окремо підрядні й головні речення при цьому залишаються програмою поза увагою. Якщо у складному реченні буде збігатися з перекладацькою пам'яттю тільки словосполучення чи одне з підрядних речень, програмою машинного перекладу такий збіг буде знехтуваний як такий, що становить менше 70% (або значно менше). Можна навіть припустити, що надто ускладнене речення, скажімо, з одним головним і трьома підрядними, майже ніколи не повториться в різних текстах, і в перекладацькій пам'яті буде лише займати обсяг, тобто, фактично, не бути корисним. Звичайно, було б добре, якби в перекладацькій пам'яті містилися і фрагменти дрібніші, ніж речення, адже речення можуть виявитися надзвичайно масштабними. Отже, пропонуємо алгоритм збагачення перекладацької пам'яті шляхом її розбиття на еквіваленти дрібніші, ніж пунктуаційно оформлені речення.

Тому, на наш погляд, одним із перспективних напрямків поліпшення систем машинного перекладу може стати збагачення перекладацької пам'яті на основі уже наявних перекладних фрагментів. Відтак, мета статті – представити в загальних рисах алгоритм програми (уже наявної), здатної автоматизувати цю роботу.

Питання розширення перекладацької пам'яті вже давно на часі. Частково цю функцію виконує термінологічна база даних, що вбудовується в більшість таких програм. У свою чергу, термінологічна база даних може бути скомпільована автоматично. Програми автоматизованої вибірки термінів (англійською – «term-extractors") у своїй більшості одномовні, однак існують і двомовні, які здатні вибирати еквіваленти термінів із паралельних (або псевдопаралельних) вирівняних текстів, тобто, фактично укладати двомовний словник. Під паралельними текстами маються на увазі оригінал і переклад. Вирівняним текстом у програмах машинного перекладу називається такий текст, який оформлений у базу даних, де проти комірки фрагмента перекладу розташований еквівалентний фрагмент оригіналу (найчастіше – речення). Не зустрічалися нам програми, що здатні здійснювати вибірку еквівалентних термінів із невирівняних текстів. Серед двомовних програм вибірки термінів, що працюють з уже укладеною перекладацькою пам'яттю, слід назвати *Arraya Term Extractor, Multi Term Extract (SDL), Promt Terminology Manager*; усі названі програми платні. Укладання двомовних термінологічних глосаріїв корисне не стільки з погляду публікації словників, скільки для збагачення комп'ютерної перекладацької пам'яті. Для публікації словника цей матеріал надто сирий: потрібна ретельна перевірка синонімів, можливих контекстів уживання. Натомість для перекладача, що працює у вузькоспеціалізованій

галузі, звичні перекладацькі еквіваленти *ad hoc* будуть надзвичайно корисні. У такий спосіб, вибрані терміни поповнюють термінологічну базу даних проекту, і при наступному запуску програма пропонуватиме для них еквіваленти автоматично.

Розробники програм машинного перекладу подбали про дрібні фрагменти теж: такі популярні знаряддя, як *Trados, Déjà Vu*, дозволяють інтегрувати дво- та багатомовні термінологічні бази даних. Це означає, що програма машинного перекладу вже на початковому етапі виявить у тексті оригіналу терміни й запропонує для них переклади. Позаяк загальною тенденцією термінології є наявність одно-однозначних еквівалентів, можна сподіватися, що запропоновані переклади для термінів будуть часто безпомилковими. Винятками стануть справді багатозначні терміни, не сприйняті складні (дво-, трикомпонентні), а також нетермінологічна лексика, які в іншій галузі могли б виявитися терміном.

На наш погляд, цей, безумовно, конструктивний підхід дуже перспективний. Однак при ньому нехтується проміжна ланка: виходячи з теорії закономірних відповідників (Я.Й. Рецкер), закономірними відповідниками, еквівалентами при перекладі можуть виступати не лише терміни. Чимало клішованих виразів, словосполучень, зв'язкових та вставних слів і виразів, конструкцій, ідіом можуть мати доволі сталі відповідники, і при цьому не потрапляти до перекладацької пам'яті. Певна річ, ідеться не про те, щоб завантажувати до перекладацької пам'яті словник ідіом чи кліше, хоча і це може певним чином поліпшити якість перекладів, оскільки навіть суто навмання деякі завантажені вирази можуть-таки трапитись у тексті. Але оскільки основна ідея перекладацької пам'яті грунтується на наявності збігів при перекладі великої кількості однотипних текстів чітко визначеної тематики, жанру, стилю, навіть автора й адресата, то й відповідно словники закономірних відповідників необхідно укладати *ad hoc*, для заданої категорії текстів. Безперечно, для укладання комп'ютерного словника "на один проект" застосовуватиметься відповідно спрощений підхід. Адже укладання словника лише заради одного замовлення виявляється нераціональним з погляду використання людських ресурсів і, відповідно, у фінансовому плані. Таким чином, необхідна розробка програми, яка буде здійснювати вибірку закономірних двомовних відповідників для текстів заданого типу, теми, стилю, жанру, автора, замовника тощо. Традиційні програми вибірки термінів для цього не підходять, адже закономірні відповідники – не лише терміни. Автор статті дозволив собі здійснити спробу розробити власну програму, оскільки в будь-якому разі актуальним видається опробування нового алгоритму; поза тим, відомі програми надто коштовні й, відтак, недоступні для перекладачів, викладачів перекладу.

Програми вибірки закономірних відповідників можуть бути прив'язані або не прив'язані до конкретних мов або ні. У першому випадку, на наш погляд, алгоритм обробки перекладацької пам'яті виявиться надто складним, і наразі ми його не використовували, що залишається завданням на перспективу. Прив'язка до певних мовних пар, на перший погляд, звужує можливість використання. У той самий час, цей недолік доволі незначний, якщо врахувати, що певний перекладач рідко працює з більш, ніж трьома-п'ятьма мовами. Може здаватися, що розробка програми потребуватиме введення величезної кількості даних щодо кожної мови або мовної

пари. Однак це не так: у пропонованій програмі введення близько ста закономірних відповідників дозволяє успішно виконувати поставлене завдання. Відтак, наразі програма працює з десятьма мовними парами: англо-українська, іспансько-українська, французько-українська, італійсько-українська, польсько-українська та *vice versa*.

Алгоритм програми передбачає два варіанти: автоматична вибірка та напівавтоматична.

Автоматичний алгоритм працює з вирівняними текстами (а в перспективі – і з невирівняними). У разі виявлення у реченні оригіналу й перекладу еквівалента, що занесений у базу даних, програма використовує його як орієнтир для розбиття речення на фрагменти. Виявлені фрагменти оригінального й перекладного речення зіставляються програмою й після перевірки деякими фільтрами видаються як готові еквіваленти.

Напівавтоматичний алгоритм спочатку пропонує користувачеві вибрані програмою еквіваленті, що здійснюється на основі аналізу повторюваних збігів. Після підтвердження користувачем правильності обраних варіантів і відхилення неправильних програма поповнює запас міжмовних закономірних відповідників, які додатково використовує для розбиття речень на еквіваленти.

У результаті роботи програми точність знайдених еквівалентів становить від 90% і вище. Хоча програма призначається для роботи з нехудожніми текстами, вказаний алгоритм виявляється продуктивним і для опрацювання перекладів художньої прози. Незначні відхилення між оригіналом і перекладом трапляються тоді, коли в перекладі речення відбулися перестановки. Програма може схибити, наприклад, у зв'язку з тим, що в іспанській, французькій, італійській, польській мовах прикметник уживається найчастіше в постпозиції по відношенню до іменника. Втім, фільтри дозволяють позбутися значної кількості негативного матеріалу (так званого «сміття»), і кількість негативного матеріалу мінімальна. Для прикладу тут наводимо вибірку еквівалентів з англо-української пари текстів другої статті Статуту ООН, які умовно називаємо квазіпаралельними (ті, що не є перекладом один одного, однак перекладені з одного оригіналу). З огляду на обмеженість обсягу, обрано фільтр довжини еквівалента – не більше 70 символів, тому фрагменти, що могли б мати масштаб речення, в ілюстрації відсутні:

1. action – діях
2. action it takes in – діях що
3. action it takes in accordance with the present Charter – діях, що вживаються нею відповідно до даного Статуту
4. against which the United Nations is taking preventive – проти якої Організація Об'єднаних Націй вживає дії превентивного
5. and justice are not endangered – і справедливість
6. and security – та безпеку –
7. and security and justice are not endangered – і справедливість
8. and shall refrain – і утримуються
9. and shall refrain from – і утримуються від
10. and shall refrain from giving – і утримуються від надання

11. and shall refrain from giving assistance to any – і утримуються від надання допомоги будь-якій
12. assistance – допомоги
13. assistance to any – допомоги будь-якій
14. be necessary – виявитися необхідним
15. but – однак
16. ensure that – забезпечує щоб
17. for the maintenance of international peace and security – для підтримки міжнародного миру й безпеки
18. from – від
19. from giving – від надання
20. from giving assistance to any – від надання допомоги будь-якій
21. giving – надання
22. giving assistance to any – надання допомоги будь-якій
23. in any – у всіх
24. in any action it takes in – у всіх діях що
25. international peace – міжнародний мир
26. international peace and security – міжнародний мир та безпеку
27. international relations from the threat – застосування як проти територіальної недоторканності
28. it takes in accordance with the present Charter – що вживаються нею відповідно до даного Статуту
29. justice are not endangered – справедливість
30. manner that international peace – таким чином, щоб не наражати на загрозу міжнародний мир
31. manner that international peace and security – таким чином, щоб не наражати на загрозу міжнародний мир та безпеку
32. or in any – так і якимось
33. or political independence of any state or in any – або політичної незалежності будь-якої держави так і якимось
34. peace – мир
35. peace and security – мир та безпеку
36. political independence of any – політичної незалежності будь-якої
37. political independence of any state or in any – політичної незалежності будь-якої держави так і якимось
38. security – безпеку
39. shall refrain – утримуються
40. shall refrain from – утримуються від
41. shall refrain from giving – утримуються від надання
42. shall refrain from giving assistance to any – утримуються від надання допомоги будь-якій
43. state against which the United Nations is taking preventive – державі проти якої Організація Об'єднаних Націй вживає дії превентивного

44. state or in any – держави так і якимось

45. that – щоб

46. the maintenance of international peace and security – підтримки міжнародного миру й безпеки

47. their international relations from the threat – її застосування як проти територіальної недоторканності

48. to any – будь-якій

49. with the present Charter – відповідно до даного Статуту

50. with the present Charter and shall refrain – відповідно до даного Статуту і утримуються

51. with the present Charter and shall refrain from – відповідно до даного Статуту і утримуються від

52. with the present Charter and shall refrain from giving – відповідно до даного Статуту і утримуються від надання [5], [6].

Еквіваленти 1, 14, 15, 21, 34, 35, 38 цілком можуть поповнити комп'ютерний словник однослівних відповідників. Вони не є термінами, однак їх доречно розглядати як закономірні відповідники нейтральної лексики. Еквіваленти 3, 17, 25, 26, 46, 49 можуть виступати закономірними відповідниками словосполучень, що цілком імовірно можуть трапитись у інших документах.

У сучасних програмах зазначені приклади, корисні як потенційні еквіваленти для перекладу, не потрапили б ані до перекладацької пам'яті, оскільки не є завершеними реченнями, ані до термінологічної бази даних, позаяк вони не є термінами. Таким чином, потенціал подібних програм у збагаченні перекладацької пам'яті доволі високий. Переваги програми в тому, що, на відміну від традиційного вирівнювання тексту, яке відбувається часто вручну, вибірка слів і словосполучень відбувається автоматично, а роль користувача – тільки підтвердити або відхилити запропонований варіант. З поданого ілюстративного прикладу видно, що точність знайдених еквівалентів перевищує 90%, а приблизно 25% з автоматично вибраних еквівалентів можуть бути використані як потенційні еквіваленти для подальших перекладів.

Першочерговим завданням у перспективі вважаємо перевірку результатів використання програми в експерименті. Менш терміновим, але не менш важливим завданням на майбутнє вважаємо укладання програми, яка працюватиме не лише з вирівняними текстами.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Електронний словник мови Тараса Шевченка. –[Електронний ресурс] – Режим доступу від 30.09.2016:http://www.mova.info/Page.aspx?l1=220 2. Словник порівнянь Юрія Андруховича та Оксани Забужко – [Електронний ресурс] – Режим доступу від 30.09.2016:http://www.mova.info/porivn.aspx?l1=643. 3. Listado de Frecuencias del Corpus de Referencia del Español Actual – [Електронний ресурс] – Режим доступу від 30.09.2016:http://corpus.rae.es/lfrecuencias.html.

ДЖЕРЕЛА ІЛЮСТРАТИВНОГО МАТЕРІАЛУ

4. Carta de la Organización de las Naciones Unidas – [Електронний ресурс] – Режим доступу від 30.09.2016: http://www.un.org/es/carta-de-las-naciones-unidas/index.html5. Статут Організації Об'єднаних Націй. – [Електронний ресурс] – Режим доступу від 30.09.2016: http://www.un.org.ua/images/UN_Charter_Ukrainian.pdf. 6. Charter of the United Nations. –[Електронний ресурс] – Режим доступу від 30.09.2016: http://www.un.org/en/charter-united-nations/.

Фокин С.Б., к. филол. н., доц.,
Институт филологии КНУ имени Тараса Шевченко, Киев

СПОСОБЫ АВТОМАТИЗИРОВАННОГО ОБОГАЩЕНИЯ КОМПЬЮТЕРНОЙ ПЕРЕВОДЧЕСКОЙ ПАМЯТИ

В статье рассматривается алгоритм разработанной автором компьютерной программы, позволяющей обогащать переводческую компьютерную память, а, следовательно, совершенствовать качество машинного перевода; обогащение переводческой памяти предусматривает увеличение количества потенциальных эквивалентов предложений, словосочетаний, нетерминологических лексических единиц. В то же время большинство современных программ направлены на составление переводческой памяти из целых предложений, а также терминологической базы данных; при этом промежуточное звено (части сложных предложений, нетерминологические устойчивые словосочетания) остаются незадействованными в работе программы машинного перевода.

Ключевые слова: *автоматическая выборка, термины, машинный перевод, алгоритм, закономерные соответствия.*

Fokin S.B., PhD., Associate Professor
Taras Shevchenko National University of Kyiv

AUTOMATIC MEANS OF ENHANCING COMPUTER'S TRANSLATION MEMORY

The article deals with the algorithm of an author's program, which is intended for enriching the translation memory, hence, for improving the quality of machine translation; enriching contemplates an augmentation of potential equivalents of sentences, clauses, phrases and non-terminological lexical units. At the same time, many programs aim to compile a translation memory consisting of whole sentences and a term base, the intermediate components (parts of complex sentences, non-terminological words and word combinations) remain unimplemented in machine translation programs.

Kew words: *automatic extraction, terms, machine assisted translation, algorithm, regular correspondences.*