УДК 81’33:81’255

Юрченко О.М.
(Харків, Україна)

СЕГМЕНТАЦІЯ ТЕКСТУ ПЕРЕКЛАДУ ЗА СЛОВОСПОЛУЧЕННЯМИ У COMPUTER-AIDED TRANSLATION СИСТЕМАХ

У статті запропонована система сегментації тексту перекладу за принципом виділення словосполучень у якості одиниці перекладу та представлено алгоритм роботи програмного забезпечення для реалізації виділення словосполук у тексті та їх наступного перекладу.

Ключові слова: *сегментація тексту, автоматизовані системи перекладу, дієслівні словосполучення, одиниця перекладу.*

В статье предложена система сегментации текста по принципу выделения словосочетаний в качестве единицы перевода и представлен алгоритм работы программного обеспечения, предназначенный для выделения словосочетаний в тексте и их последующего перевода.

Ключевые слова: *сегментация текста, автоматизированные системы перевода, глагольные словосочетания, единица перевода.*

The article offers the text segmentation system by word-combinations like the unit of translation and an algorithm of software for performing text segmentation and further translation of obtained segments.

Key words: *text segmentation, cat-tools, automated translation, computer-aided translation system, verb patterns, translation unit.*

За останні роки характер роботи перекладача і вимоги до нього істотно змінилися. У першу чергу зміни торкнулися перекладу науково-технічної, офіційної та ділової документації. У наш час вже недостатньо просто перекласти текст, користуючись комп’ютером як друкарською машинкою. Замовник очікує від перекладача, що оформлення готового документа буде відповідати зовнішньому вигляду оригіналу настільки точно, наскільки це можливо, при цьому задовольняти прийнятим у даній країні стандартам. Терміни виконання таких замовлень зменшуються, тому перекладачеві, який не користується допоміжними засобами автоматичної обробки мовної інформації, важко вкладатися в установлені часові рамки. Цих жорстких, часто суперечливих умов можливо дотримуватись лише у тому випадку, якщо перекладач не тільки досконало володіє рідною та іноземною мовами та глибоко вивчив обрану ним предметну область, але й впевнено орієнтується в сучасних комп’ютерних технологіях [1].

Ключовою для перекладача текстів технічної направленості є технологія Translation Memory (ТМ − пам’ять перекладів), що широко використовується в системах автоматизованого перекладу (АП чи CAT від англ. Computer-Aided Translation). Автоматизований переклад – переклад текстів з використанням комп’ютерних технологій. Від машинного перекладу (МП) він відрізняється тим, що весь процес перекладу здійснюється людиною,

комп'ютер лише допомагає їй виробити готовий текст або за менший час, або з кращою якістю.

Автоматизований переклад передбачає такі форми взаємодії людини з комп'ютером:

• Частково автоматизований переклад: наприклад, використання перекладачем-людиною комп'ютерних словників.

• Системи з розподілом праці: комп'ютер, може перекладати тільки фрази жорстко заданої структури (але робить це так, щоб виправляти за ним не було потрібно), а все що не вкладається в схему – перекладається людиною.

Ідея АП з'явилася з моменту появи комп'ютерів: перекладачі завжди виступали проти стандартної у ті роки концепції МП, на яку було спрямовано більшість досліджень в області комп'ютерної лінгвістики, але підтримували використання комп'ютерів для допомоги перекладачам. У 1960-ті роки Європейське об'єднання вугілля і сталі (попередник сучасного Євросоюзу) стало створювати термінологічні бази даних під загальною назвою Eurodicautom. У Радянському Союзі для розробки баз такого роду був створений ВІНІТІ.

У сучасній формі ідея АП була розвинена в статті Мартіна Кея (1980 р.), який висунув наступну тезу: «By taking over what is mechanical and routine, it (computer) frees human beings for what is essentially human» («Комп'ютер бере на себе рутинні операції і звільняє людину для операцій, що вимагають людського мислення») [2].

Найбільш розвиненими та розповсюдженими на наш час CAT-системи вважаються наступні: SDL Традос, Deja Vu, Wordfast, (що не є повноцінною CAT-системою, а представляється у якості надбудови до MS Word). Все ці системи автоматизованого перекладу ґрунтуються на пам'яті перекладів (ПП, англ. Translation memory (TM) − так званий «Накопичувач перекладів»), тобто базі даних (БД), що містить набір фраз із раніше перекладених текстів [3]. Один запис в такій базі даних відповідає «одиниці перекладу» (англ. translation unit), за яку зазвичай приймається одне речення (рідше − частина складносурядного речення, або абзац). Якщо чергове речення вихідного тексту в точності збігається з реченням, що зберігається в базі (точна відповідність, англ. Exact match), воно може бути автоматично підставлено в переклад. Нове речення може також злегка відрізнятися від того, що зберігається в базі (неточна відповідність, англ. Fuzzy match). Таке речення може бути також підставлено в переклад, але перекладач повинен буде внести необхідні зміни. Бази даних пам'яті перекладу можуть бути складені також із вже виконаних перекладів за допомогою вбудованих функцій на самому початку роботи з програмою.

У кожній конкретній системі пам'яті перекладів дані зберігаються у своєму власному форматі (текстовий формат у Wordfast, база даних Access в Deja Vu і т.ін), але існує міжнародний стандарт TMX (англ. Translation Memory eXchange format), що заснований на XML і з яким можуть працювати практично всі системи пам'яті перекладів.

Крім прискорення процесу перекладу повторюваних фрагментів і змін, внесених до вже перекладених текстів (наприклад, нових версій програмних продуктів або змін у законодавстві), системи ПП також забезпечують однаковість перекладу термінології в однакових фрагментах, що особливо важливо при технічному перекладі. З іншого боку, якщо перекладач регулярно підставляє в свій переклад точні відповідності, витягнуті з баз перекладів, без контролю їх використання в новому контексті, якість перекладеного тексту може погіршитися.

Перевагами перекладання за допомогою CAT-засобів є:

1. Забезпечення однаковості перекладу, що позитивно позначається на його якості;

2. Прискорення темпу робіт з перекладання за рахунок можливості не перекладати однакові фрагменти тексту двічі. В результаті, скорочуються терміни, необхідні на переклад;

3. Можливість внесення зміни, доповнень та зауважень замовника по всій базі перекладів, що дозволяє миттєво виправити неточності в уже перекладених сегментах і уникнути нової появи подібних помилок.

4. У разі якщо вихідний документ наданий в одному з розповсюджених форматів (напр., Microsoft Word (.doc); Microsoft Excel (.xls); Microsoft PowerPoint (.ppt); документи QuarkXPress; документи Adobe InDesign; документи Adobe Framemaker; документи Adobe Pagemaker; HTML-сторінки (.html, .htm); файли довідки MS Windows (.chm); розширювана мова розмітки (.Xml) та ін..), переклад здійснюється без порушення структури документа. Фактично необхідно лише відкоригувати текстові фрагменти для усунення невідповідності обсягів тексту оригіналу і перекладу.

5. CAT-засоби дозволяють знизити загальну вартість перекладу для замовника. На відміну від звичайних розрахунків на переклад, переклад за допомогою CAT-засобів тарифікується по кількості слів з урахуванням наступних параметрів:

• кількість однакових (повторюваних) сегментів;

• кількість сегментів, співпадаючих із пам'яттю перекладів у відсотковому співвідношенні.

Таким чином, стає зрозумілим, що одним з основних принципів роботи CAT-систем є сегментація тексту, тобто лінійне членування мовного потоку на складові відрізки − сегменти, співвідносні з певними одиницями мови: значущими реченнями, словосполученнями, словами та навіть морфемами.

Оскільки одиницею будь-якого конкретного або абстрактного об'єкта прийнято називати його елементарну частину, що зберігає всі характеристики цілого, то одиниця перекладу, що виділяється, повинна збігатися з одиницею мовлення вихідного (перекладеного) тексту, сенс чи інформацію якої перекладач повинен зрозуміти, еквівалент, відповідність якої він повинен підібрати . Тобто одиницю перекладу слід шукати в початковому тексті. Вона являє собою одиницю мови, що вимагає окремого рішення на переклад. Інакше кажучи, одиниця перекладу – це така одиниця в початковому тексті, яка повинна бути виділена і якій можливо підшукати еквівалент в тексті перекладу, але складові частини, якої окремо не мають відповідності в тексті перекладу як текстові одиниці.

Саме поняття «одиниця перекладу» певною мірою умовно, бо не є величиною постійною. Враховуючи асиметрію пар мов, що включаються в процес перекладу, в якості основних одиниць перекладу в процесі сегментації вихідного тексту можуть виступати слово, словосполучення, речення, надфразові єдності.

Слід акцентувати увагу на тому, що практично всі системи автоматизованого перекладу членують текст на речення. Але речення, на наш погляд, є досить крупною одиницею перекладу для CAT-систем, оскільки повний збіг подібних сегментів-речень можливий лише у формальних елементах оформлення документа (шапках, колонтитулах і т.п.) [4]. Тому нашою метою стала розробка такої автоматизованої системи перекладу, яка б сегментувала текст, беручи за одиницю членування словосполучення, яке з одного боку,

є найпростішою синтаксичною одиницею мови, з іншого, є одиницею мови, що несе у собі певне граматичне та семантичне значення.

У якості матеріалу для дослідження було обрано дієслівні словосполучення у перекладі з англійської мови на російську. Одним з найбільш розповсюджених типів дієслівних словосполучень в англійські мові є Verb Patterns (VP), які будуються за наступними моделями [5]:

1) Дієслово + інфінітив (V + to + V);
2) Дієслово + ing-форма (V + V-ing);
3) Дієслово + інфінітив або ing-форма;
4) Дієслово + об'єкт + інфінітив (без to);
5) Дієслово + об'єкт + ing-форма (герундій).

Формальному описанню наведених моделей дієслівного керування сприяє також той факт, що перелік перших дієслів даних VP є сталим. Семантика таких конструкцій буде залежати від другої дієслівної форми словосполучення, а форма підпорядковується вимогам мови перекладу та стилістики тексту. Наприклад, у російській мові VP можуть перекладатися як дієслівними словосполученнями дійсного чи пасивного стану, словосполученнями дієслово + іменник і навіть складними реченнями:

1) **You are not *allowed to wear* shorts in office.**
*Рос. Вам не **позволено носить** шорты в офисе.*
*Рос. Тебе не **разрешали надевать** шорты на работу.*
*Рос. Тебе нельзя **приходить на работу** в шортах.*
2) **Did you *enjoy visiting* the zoo?**
*Рос. Тебе **понравился поход** в зоопарк?*
*Рос. Вы **получили удовольствие** от посещения зоопарка?*
*Рос. Ты **доволен походом** в зоопарк?*
*Вы **остались довольны тем, как сходили** в зоопарк?*

3) **She *likes speaking* loudly.**
*Рос. Она **любит** громко **говорить**.*
*Рос. Ей **нравится разговаривать** громко.*

We *like to watch* movies.
*Рос. Мы **любим смотреть** фильмы.*
*Рос. Нам **нравится смотреть** кино.*

4) **They *helped us to write* the book.**
*Рос. Они **помогли нам написать** книгу.*
*Рос. Они **помогали нам писать** книгу.*
*Рос. Они **оказали нам помощь** в написании книги.*

Результатом дослідження є алгоритм роботи програмного забезпечення, за яким виконується членування тексту на словосполучення та пропонуються варіанти перекладу виділених словосполучень, як це схематично представлено на рис. 1.

Рисунок 1 − Схема роботи САТ-системи із сегментацію за словосполученнями.

Таким чином, згідно з запропонованим нами алгоритмом:

• по-перше, буде здійснюватися обробка та сегментація вихідного тексту;

• по-друге, провадитися порівняння отриманих сегментів із пам'яттю перекладів;

• в результаті можливо змінити переклад запропонованих сегментів; додати переклад до нових сегментів чи відкинути зайві сполуки, що не мають відповідного змісту.

Проведене дослідження показало, що сегментація тексту перекладу у САТ-системах за принципом виділення словосполучень є більш ефективною. Автором запропонований алгоритм роботи автоматизованої системи перекладу із сегментацію за дієслівними словосполученнями типу англійських Verb Pattens. У перспективі подальшого дослідження планується програмно реалізувати запропонований алгоритм та розглянути інші типи словосполучень та інші мовні пари.

ЛІТЕРАТУРА

1. Алимов В.В. Теория перевода. Перевод в сфере профессиональной коммуникации. – М.: Едиториал УРСС, 2005. – 80 с.
2. Кей Мартін. Домашня сторінка в університеті Саарланда. − http://www.stanford.edu/~mjkay/
3. Бархударов Л.С. Язык и перевод. Вопросы общей и частной теории перевода. – М.: Международные отношения, 1975. – 42 с.
4. John Hutchins. The origins of the translator's workstation. – NY: Kluwer Academic Publishers, 1998. – 287 – 307 p.
5. Казакова Т.А. Практические основы перевода. English-Russian. – СПб.: Лениздат; 2003. – 35 с.