УДК 81’25(075.8)

МАШИННИЙ ПЕРЕКЛАД У КОНТЕКСТІ СУЧАСНОГО НАУКОВО-ТЕХНІЧНОГО ПЕРЕКЛАДУ

А.Л. Міщенко, канд. філол. наук (Кіровоград)

У статті описується модель технічної взаємодії пам’яті перекладу, машинного перекладу та постредагування, що дозволяє забезпечувати оптимальну якість перекладу й заощаджувати час та кошти, але вимагає значних лінгвістичних ресурсів у формі цифрових даних для “тренування” системи.

Ключові слова: переклад, перекладач, пам’ять перекладу, машинний переклад, редагування, якість перекладу, заощадження часу й коштів на переклад, лінгвістичні ресурси у формі цифрових даних.

Мищенко А.Л. Машинный перевод в контексте современного научно-технического перевода. В статье описывается модель технического взаимодействия памяти перевода, систем машинного перевода и постредактирования, что позволяет обеспечивать оптимальное качество перевода, сокращает стоимость и время перевода, но требует наличия значительных лингвистических ресурсов в электронной форме для “обучения” системы.

Ключевые слова: перевод, переводчик, память перевода, машинный перевод, редактирование, качество перевода, стоимость и время перевода, лингвистические ресурсы в электронной форме.

Mishchenko A. Machine Translation within the Context of Modern Translation in Science and Technology. The paper offers a model for the technical correlation of translation memory, machine translation systems and editing that enables to achieve a high-quality translation, lower levels of translation cost and time, but requires considerable linguistic electronic resources to ‘teach’ the system.

Key words: translation, translator, translation memory, machine translation, editing, high-quality translation, translation cost and time, linguistic electronic resources.

Величезні обсяги перекладу, велика кількість мов перекладу й жорсткі умови цейтноту, з одного боку, та спеціалізація перекладу для вузької предметної галузі, контрольовані мови, уніфіковані лінгвістичні ресурси, з іншого боку, стимулювали черговий ренесанс машинного перекладу. Тому об’єктом вивчення цієї статті є системи машинного перекладу, а предметом аналізу обрано специфіку їхньої інтеграції в системи пам’яті перекладу.

Сьогодні системи машинного перекладу, які базуються на правилах, статистичні й гібридні системи машинного перекладу інтегруються в системи пам’яті перекладу, а “розумна” синергія лінгвістичних ресурсів, технологій і прикладних програм реалізується в оптимальній взаємодії перекладача, систем пам’яті перекладу і машинного перекладу. Це зумовлює актуальність обраної теми. Тому метою цієї статті обрано висвітлення сучасної моделі перекладу, основу якої складають програмні продукти й лінгвістичні ресурси. Реалізація мети передбачає вирішення таких завдань: висвітлити процес інтеграції систем машинного перекладу в системи пам’яті перекладу, а також проілюструвати функціонування інтегрованої моделі перекладу.

Системи машинного перекладу диференціюються на ті, що базуються на правилах, статистичні та гібриді. Ефективність систем машинного перекладу, які базуються на правилах (RBMT: regelbasierte maschinelle Translation), визначається якістю двомовних словників та точністю заданих правил, а їхнє створення потребує тривалої кропіткої роботи. Раніше такі системи розроблялись тільки

на замовлення "великих" організацій із залученням міжнародних робочих груп дослідників. На сучасному етапі доступність ліцензій дозволяє середнім і малим підприємствам адаптувати їх до перекладу власного контенту.

Еволюцію систем машинного перекладу, які базуються на правилах, схематично, але наочно ілюструє модель трикутника, запропонована відомим французьким математиком і програмістом Бернардом Вокуа (Bernard Vauquois).

Рис. 1. Модель машинного перекладу (Bernard Vauquois)

Як це видно з Рис. 1, перші системи машинного перекладу створювалась для конкретних пар мов і базувались на складних процесах моделювання мови, основу яких становили методи аналізу, трансферу, синтезу й інтерлінгви. Перші системи працювали на основі прямого методу заміни слів мови оригіналу словами мови перекладу. Системи другого покоління аналізували структури мови оригіналу, а потім на основі трансферу синтезувала їх в еквівалентні структури мови оригіналу. Цей метод спочатку був застосований для синтаксичного, а вже потім для семантичного рівня мови. Крім того, робилися спроби створити системи машинного перекладу на основі формальної мови-посередника – інтерлінгви. Ця концепція передбачала "переклад" літерального значення речення мовою оригіналу формальною мовою посередником з його наступною генерацією мовою перекладу. У якості інтерлінгви використовували латину, есперанто, мову американських індійців – аімара, але на основі інтерлінгви не було створено жодної ефективної системи машинного перекладу. Таким чином, ці системи базувались на різних рівнях лінгвістичного опрацювання мовної пари, які поетапно додавались у процесі еволюції галузі, а саме:

1) морфологічному: лематизація лексичних одиниць, пошук лексичних одиниць у словнику, аналіз морфем, розпізнавання контекстного граматичного класу лексичних одиниць, відмінків, флексій тощо;

2) синтаксичному: розпізнавання типів синтаксичних структур, реляційних зв'язків між окремими елементами синтаксичної структури тощо;

3) семантичному: виокремлення лексичного значення багатозначних лексичних одиниць та афіксів, визначення їхньої семантичної функції, синтез їхньої синтаксичної однозначності на основі семантичного аналізу.

Проте релевантність мовних рівнів визначається структурою мови. Звідси випливає, що для конкретної пари мов лінгвістичне опрацювання цих рівнів подекуди асинхронне, напр.: для пари мов китайська – німецька. Це пояснюється тим, що морфологічна структура китайської мови надзвичайно проста, тому морфологічний аналіз для неї не релевантний. Проте складна морфологія німецької мови вимагає обов'язкового морфологічного аналізу.

Сьогодні, наприклад, фірма Delta International (http://www.dicits.com/index.php?option=com_content&view=article&id=21&Itemid=21&lang=ru) застосовує професійні системи перекладу Lucysoft, SYSTRAN, Prompt для оптимізації процесу перекладу з використанням систем пам'яті перекладу та адаптації цих систем до потреб замовника шляхом додавання у систему словників корпоративної термінології та програмування лінгвістичних правил для бі-директального перекладу текстів для таких пар мов, як: англійська – німецька, німецька – російська, англійська – російська тощо.

Сучасні системи машинного перекладу базуються переважно на статистичних або гібридних методах. В основі перших - автоматична екстракція схожих сегментів мовних пар з двомовних повнотекстових корпусів, які нараховують мільярди слововживань. Другі – створюються сьогодні на ґрунті існуючих систем машинного перекладу, що базуються на правилах, додаванням до них статистичних методів. Таким чином, "навчання" як статистичних систем машинного перекладу (ССМП), так і гібридних систем машинного перекладу (ГСМП) базується на двомовних корпусах текстів і не потребує глибокого й складного контрастивного лінгвістичного аналізу. Це дозволило суттєво знизити вартість ССМП, що, у свою чергу, стало визначальним фактором їхньої популяризації й стрімкого поширення. Так, вартість традиційних систем машинного перекладу сягала сотні тисяч доларів, а ціна сучасних статистичних систем машинного перекладу не перевищує кілька тисяч доларів. Їхня доступна ціна й популярність спонукала розробників систем пам'яті перекладу оптимізувати архітектуру своїх продуктів за рахунок можливості долучення до ССМП з робочого середовища систем пам'яті перекладу.

Так, доступ до систем машинного перекладу з пам'яті перекладу пропонують сьогодні своїм клієнтам світові лідери, які спеціалізуються на створенні програмних продуктів для галузі лінгвістичних послуг: SDL, Across, Killgray (MemoQ), Atril (DejaVu).

Функціонування алгоритмів систем машинного перекладу не залежить від комбінації мов, тому їх використовують для "навчання" нових пар мов, а сама можливість такого тренування гарантує поступове покращання якості перекладу. Проте для забезпечення ефективної організації процесу перекладу необхідно враховувати наступне: 1) специфіку контенту й можливості застосування машинного перекладу; 2) наявність двомовного корпусу для тренування системи (обсяг корпусу від одного до п'яти мільйонів слововживань); 3) потенційні можливості для контролю якості такого перекладу. Зупинимось докладніше на інтегрованій моделі системи пам'яті перекладу SDL та системи статистичного машинного перекладу Language Weaver (LW), оскільки саме вона пропонує клієнтам новаторське рішення: тренування системи ресурсами пам'яті перекладу замовника.

На відміну від систем пам'яті перекладу інших виробників, які з робочого середовища своїх програмних продуктів уможливлюють доступ до тієї чи іншої системи машинного перекладу, корпорація SDL купила фірму Language Weaver у 2009, оскільки у СМП LW об'єднані такі системні характеристики, як комерційність, можливість доступу через Інтернет, можливість навчання системи, надійність каналу транспортування документів та можливість автоматизованого контролю якості перекладених текстів. Це рішення результувало з тривалого, глибокого й всебічного вивчення тенденцій перекладацької галузі й бажання задовольняти попит на новітні й актуалізовані продукти та послуги для створення й підтримки багатомовного контенту. Згідно з результатами дослідження SDL, 90% сучасних фірм тільки частково замовляють переклад контенту через високу вартість пе-

рекладацьких послуг. Але інші цифри вказують на те, що клієнтам SDL, які інтегрували у свої системи пам'яті перекладу машинний переклад, вдається заощаджуватий до 30% – 50% коштів на перекладі й скорочувати терміни його виконання вдвоє (Haag, Martina / Engenhart, Verena 2012). Ці цифри яскраво свідчать про те, що в майбутньому машинний переклад, інтегрований в системи пам'яті перекладу, обіцяє стати ефективним засобом скорочення витрат на переклад та дієвим інструментом розширення власної присутності на глобальному ринку для тих фірм, які практикують глобальний інформаційний менеджмент (Global Information Management). І хоча якість машинного перекладу значно поступається перекладу, зробленому перекладачем, розумна синергія перекладача, менеджменту проектами, систем пам'яті перекладу і машинного перекладу дозволяє максимально оптимізувати процес перекладу.

Фірму Language Weaver засновували Д. Марку (Daniel Marcu) та К. Найт (Kavin Knight) у 2002 р. в Лос-Анжелесі. Її мета полягала у створенні комерційної системи статистичного машинного перекладу для сервісних служб в Інтернеті, державних прес-служб та розвідувальних служб, а також для соціальних мереж. Концепцію Language Weaver президент та її генеральний директор Марк Теплінг (Mark Tapling) представив такими словами: "У той час, як перекладач Google започаткував стандарт перекладу для користувачів, ми з'ясували, що більшість фірм та підприємств хочуть мати власну технологію автоматизованого перекладу. <…> Комбінування нашої технології, технології перекладу SDL та технології менеджменту контентом дає їм унікальну можливість розширення й поглиблення контактів з існуючими та потенційними клієнтами, і це чудово узгоджується з концепцією SDL щодо глобального інформаційного менеджменту" [3].

На сучасному етапі система використовується для перекладу 24 двонаправлених пар мов (bidirektional), кількість яких зростатиме в майбутньому. Серед мов виокремлюються західноєвропейські мови, східноєвропейські мови, азійські мови, мови країн Африки та Близького Сходу.

Для перекладу в LW використовуються статистичні методи криптографії й алгоритми машинного навчання. Головний модуль Language Weaver – декодер, який здійснює керування процесом перекладу загалом, використовуючи для цього інші модулі програми, а саме: модель мови, спеціалізовані програми перекладу, спеціалізовані словники, бази даних, реєстри слів тощо.

Модуль, який відповідає за навчання системи, називається кастомайзер (Customizer). Language Weaver "навчається" безпосередньо з паралельних текстів й автоматично генерує в процесі навчання статистичну "модель мови" для цієї пари мов. Адаптація системи до потреб клієнта передбачає її навчання на матеріалі специфічних для фаху, галузі чи корпоративної мови лінгвістичних ресурсів. Ці тексти автоматично вирівнюються (Alignment) спеціальною програмою LW-Aligner або будь-якою іншою програмою вирівнювання (Aligner-Tool: Giza, WinAlign та ін.) у вигляді сегментів, представлених у формі фраз, речень, текстів чи абзаців, й зберігаються у спеціалізовану програму перекладу. На потребу програма дозволяє генерувати пам'ять перекладу і спеціалізовані словники.

Ця модель гарантує адаптацію базової системи Language Weaver до потреб клієнта й дозволяє суттєво покращувати якість перекладених текстів. Навчання системи здійснюється тільки на фірмі Language Weaver і може повторюватись через певні проміжки часу з метою актуалізації системи.

У модулі "словники" зберігаються базові й специфічні для замовника номінації у формі термінів, номенклатур, символів, формул, цифрових даних, одиниць виміру тощо. Ці лінгвістичні ресурси генеруються в процесі навчання системи перекладу і специфікуються залежно від фаху, галузі чи корпоративної мови.

Тексти з Language Weaver перекладаються в режимі пакетної обробки або в режимі реального часу, а архітектура LW передбачає можливість паралельного й незалежного використання ОЗП (основного запам'ятовувального пристрою) для перекладу. Система підтримує різні формати документів, серед яких, напр.: txt, html, tmx, xliff, pdf, odf, doc.

Процес перекладу загалом організований таким чином: необхідні документи готуються до перекладу в спеціальній субсистемі перекладу. У процесі підготовки визначається тип документів, вони перевіряються на наявність дублікатів, визначається комбінація мов, документи “очищуються” (виправляються всі можливі помилки: орфографічні, граматичні, стилістичні, логічні, змістовні тощо) і для них визначаються пріоритети. На етапі “перевірка версії” система перевіряється на застосування актуалізованої версії спеціалізованої програми перекладу або іншої визначеної для перекладу спеціалізованої програми. Після цього документи перекладаються з використанням вибраної специфічної програми перекладу. У процесі перекладу екстрагуються одиниці номінацій з унікальними назвами (Entity Names) на кшталт корпоративної лексики у формі номенклатур, термінів, номінацій продуктів, брендів і торгових марок та інших специфічних реалій і додаються до словника у формі реєстрових одиниць або словникових статей. Актуалізований варіант словника фіксується у метаданих, а перекладена версія документів зберігається. Через певні проміжки часу базовий корпус актуалізується за рахунок додавання до нього нових перекладів, які покращують якість перекладу. Процес перекладу ілюструє Рис. 2:

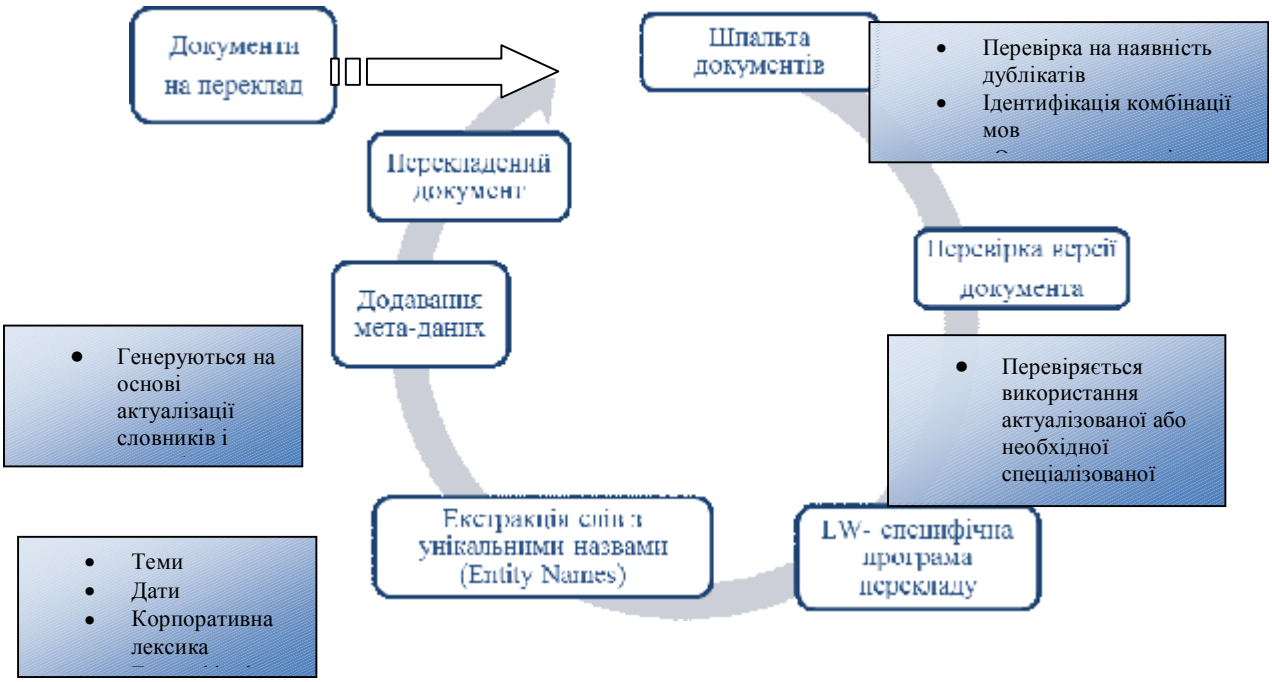

Рис. 2. Архітектура модуля перекладу в LW

.

Як зазначалося вище, Language Weaver інтегрується в продукти SDL і вможливлює адаптивний до потреб клієнта машинний переклад. Ці програмні продукти SDL диференціюються на продукти для: клієнтів, які мають власний контент і потребують ефективних інструментів його створення, актуалізації, менеджменту, перекладу тощо; державних та урядових організацій; фрилансерів.

Для власників контенту пропонується два продукти: SDL BeGlobal і SDL GlobalConnect.

SDL BeGlobal створено спеціально для транснаціональних корпорацій, тому його мета полягає в забезпеченні мультилінгвальної комунікації із застосуванням усіх доступних в електронній формі інформаційних ресурсів.

SDL-GlobalConnect – це розширений варіант

SDL BeGlobal, який створювався передусім для менеджерів контенту. Він попередньо інтегрується в чат, електронну пошту й корпоративні бази знань і на потребу дозволяє автоматичний переклад інформаційних ресурсів безпосередньо в названих прикладних програмах.

Спеціально для урядових та інших державних організацій розроблено серверну версію програми під назвою SDL Language Weaver Enterprise Translation Server, яка інтегрується в локальну систему перекладу організації в якості одного з її сегментів і дозволяє забезпечити надійний захист інформації.

Спеціально для фрилансерів і студентів створено продукт під назвою SDL Easy Translator, який являє собою своєрідного помічника комп'ютера "все в одному", здатного перекладати будь-яку текстову інформацію з будь-якого документа. Цей продукт, окрім того, дозволяє перекладати зміст популярних чат-сервісів у режимі реального часу 35 мовами.

Навчання системи з метою її адаптації до потреб клієнта триває 4 – 6 тижнів, а процес навчання схематично ілюструє Рис. 3:

Рис. 3. Language Weaver - тренування системи [2] (переклад наш – А.М.)

На етапі "навчання системи" базова версія LW розширюється за рахунок додавання галузевого або корпоративного двомовного корпусу у формі пам'яті перекладу та термінологічних баз даних. Після навчання модуль машинного перекладу інтегрується в систему перекладу й використовується в якості одного з її функціональних модулів. Якщо якість перекладених текстів недостатня, процес навчання системи повторюється.

Якість текстів автоматично перевіряється алгоритмом, який розроблено в співпраці з клієнтами в процесі адаптивного навчання системи й називається TrustScore (дослівно "оцінювання довіри"). Прогнозована якість перекладу оцінюється автоматично за п'ятибальною шкалою. Ця система контролю якості інтегрується в загальну систему менеджменту проектами й інтерпретується наступним чином. Якщо переклад оцінено в 1–2 бали

за шкалою TrustScore, він відхиляється як непридатний для використання. Переклад, який набрав 2–3 бали, надсилається перекладачеві на доопрацювання. Переклад, який оцінено у 4–5 балів, уважається придатним для репрезентації марки, але диференціюються на: 1) умовно-придатний: за наявності окремих помилок, але за певних обставин він може використовуватись для репрезентації продукту; 2) придатний для репрезентації продукту.

Таким чином, алгоритм TrustScore перевіряє якість машинного перекладу в автоматичному режимі. Схему інтеграції й організацію процесу перекладу загалом ілюструє Рис. 4:

Рис. 4. Інтегрована система перекладу: SDL-LW [2] (переклад наш – А.М.)

"Навчена" система машинного перекладу інтегрується в пам'ять перекладу SDL, у якій організація процесу перекладу функціонує наступним чином: тексти на переклад спочатку перекладаються за допомогою функції попереднього перекладу (Pre-Translation) із залученням ресурсів пам'яті перекладу (ПП). Якщо для окремих сегментів тексту повні або часткові відповідники у ПП не знайдені, вони відправляються на машинний переклад, результат якого автоматично оцінюється TrustScore і після цього, залежно від результату, текст перекладу відхиляється, відправляється на доопрацювання перекладачеві або публікується. Придатний для репрезентації марки переклад зберігається в пам'ять перекладу, а розширена ПП через певні проміжку часу використовуються для актуалізації клієнтської спеціалізованої системи перекладу з метою покращання якості МП.

Як зазначалося вище, для машинного перекладу важливо, які вимоги висуваються до якості перекладу, оскільки не всі тексти придатні для МП. Так, вимоги до якості перекладу інтерфейсу програмних продуктів, веб-контенту, реклами максимальні. Тому вони перекладаються перекладачем

з використанням ресурсів інтегрованих систем перекладу. Дещо нижчі вимоги до якості друкованих текстів зовнішньофахової комунікації, напр.: супроводжувальної технічної документації на продукти. Тому цей тип текстів перекладається зазвичай ресурсами системи адаптованого МП із наступним постредагуванням перекладача. Машинний переклад без наступного постредагування підходить для перекладу веб-контенту на кшталт повідомлень у чатах, блогах та соціальних мережах, а також для перекладу окремих текстів внутрішньофахової комунікації, напр.: листування між співробітниками, окремі тексти у Wiki тощо. Організацію процесу перекладу для різних типів текстів унаочнює Рис. 5

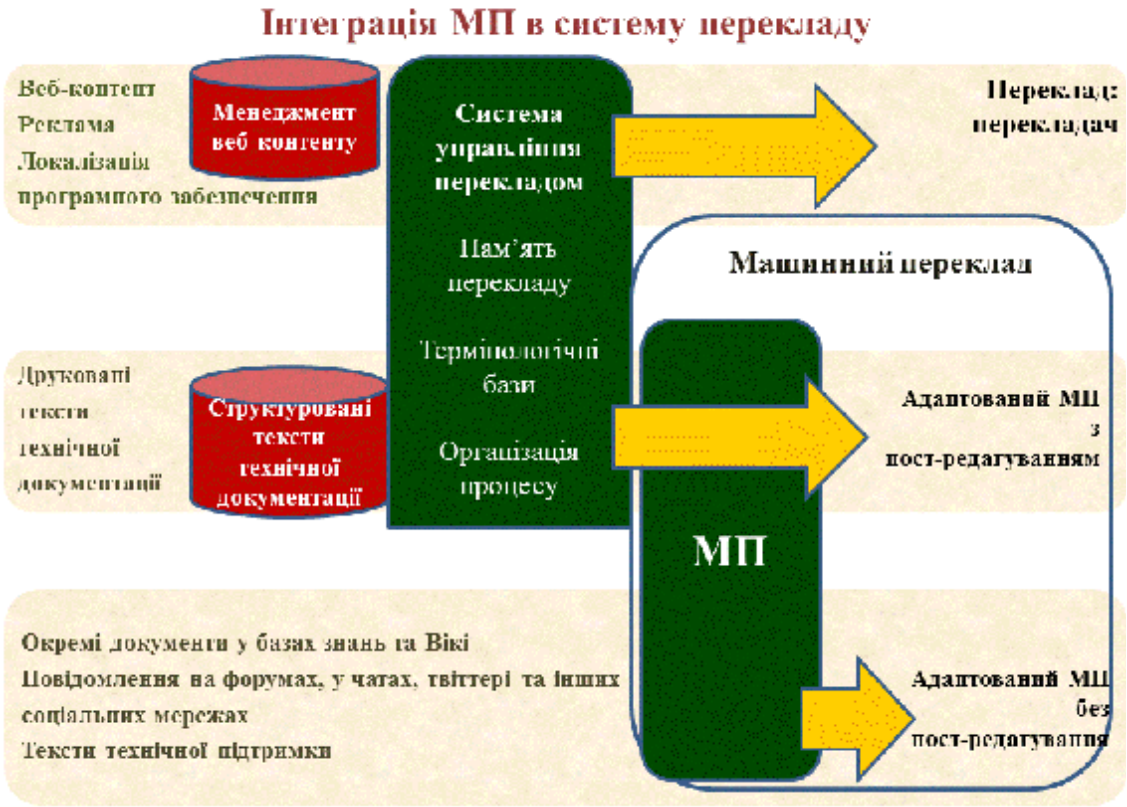

Рис. 5. Типи тексту й організація процесу перекладу

Таким чином, розумна синергія людини, машини й ефективні механізми управління дозволяють якісно покращувати переклад, мінімізувати затрати й суттєво економити час на нього. Цей "інтелектуальний" МП легко і швидко інтегрується в систему перекладу й набуває статусу її важливої складової для створення контенту, його підтримки в актуалізованому стані й управління ним як генеральної стратегії установи, а двомовні лінгвістичні ресурси дозволяють тренувати систему для перекладу нових комбінацій мов. Крім того, МП з наступним постредагуванням дозволяє швидко розширювати й актуалізувати пам'ять перекладу і таким чином оптимально використовувати існуючі лінгвістичні ресурси. Важлива перевага СМП вбачається в можливості їхньої адаптації до потреб замовника й поступового покращення якості перекладу шляхом "навчання" системи, а автоматизований контроль якості перекладу з використанням алгоритму TrustScore й зворотного зв'язку з боку замовника відіграє ключову роль для забезпечення якості МП та швидкості перекладу загалом.

Отже, комбінована модель перекладу з використанням машинного перекладу, постредагування, пам'яті перекладу, термінологічної діяльності й менеджменту процесом обіцяє стати "Соломоновим рішенням" щодо досягнення максимального ефекту мінімальними зусиллями в процесі створення й актуалізації глобального багатомовного контенту. Проте необхідною передумовою цієї моделі є технологічно складна архітектура для створення й перекладу контенту, а також двомовні лінгвістичні ресурси для "навчання" системи, тому одне з ключових завдань інформаційного суспільства полягає у створенні лінгвістичних ресурсів.

ЛІТЕРАТУРА

1. Frieling T. Language Weaver / T.Frieling. – 2009. – Access mode : http://www.slideshare.net/solution34/languageweaver [Zugriff: 25.01.2012, 17:00 MEZ]. 2. Haag M. Startklar für Maschinelle Übersetzung : Webinar / M. Haag, V. Engenhart. – Stuttgart, 03.11.2011. – Acsess Mood : http://www.youtube.com/watch?v=HxJKRz0nCyg [Zugriff: 15.01.2012, 17:30 MEZ]. 3.Tapling M. SDL Acquires Language Weaver, affirms Leadership in Machine Translation and Global Information Management. Document Transcript / M. Tapling. – 2010. – Access mode : http://www.slideshare.net/languageweaverinc/sdl-acquires-language-weaver-pr-july-15-2010. http://www.slideshare.net/TAUS/1500-sdl-language-weaver-simplifying-translation-for-the-business-world-with-automated-translation [Zugriff: 30.01.2012, 19:00 MEZ].