DOI: https://doi.org/10.20535/kpisn.2021.2.236939
УДК 007: 658.5

П.П. Маслянко*, Є.П. Сєльський
КПІ ім. Ігоря Сікорського, Київ, Україна
*corresponding author: mppdom@i.ua

МЕТОД СИСТЕМНОЇ ІНЖЕНЕРІЇ СИСТЕМ НЕЙРОННОГО МАШИННОГО ПЕРЕКЛАДУ

Проблематика. На ринку існує не так багато компаній-розробників систем машинного перекладу (СМП), продукти яких користуються попитом. Це, наприклад, "*GoogleTranslate"*, *"DeepLTranslator"*, *"ModernMT"*, *"Apertium"*, *"Trident"* тощо. Існує потреба в упорядкованих і систематизованих методів розроблення СМП, а також потрібні науково обґрунтовані методи інженерії систем нейронного машинного перекладу (СНМП), щоб якнайшвидше отримати якісний і конкурентоспроможний продукт.
Мета дослідження. Застосувати бізнес-профіль Еріксона–Пенкера для розроблення та формалізації методу системної інженерії СНМП.
Методика реалізації. Методологія системної інженерії і бізнес-профіль Еріксона–Пенкера для формалізації впорядкованого способу розроблення СНМП.
Результати дослідження. Метод розроблення СНМП на основі застосування технік системної інженерії складається з трьох основних етапів.
На першому етапі структуру СНМП моделюють як бізнес-профіль Еріксона–Пенкера, на другому – визначають множину процесів, характерну для класу систем Data Science та міжнародного стандарту CRISP-DM, а на третьому проводять верифікацію та валідацію розробленої СНМП.
Висновки. Запропоновано метод системної інженерії СНМП, що базується на модифікованому бізнес-профілі Еріксона–Пенкера представлення системи на метарівні, а також міжнародних стандартів процесів DataScience та DataMining. Досліджено ефективність застосування методу на прикладі розроблення системи двоспрямованого англійсько-українського нейронного машинного перекладу EUMT (*E*nglish-*U*krainian *M*achine *T*ranslator) і встановлено, що система EUMT щонайменше не поступається за якістю англійсько-українського перекладу популярному перекладачеві "*GoogleTranslate*".
Повний код версії системи EUMT опублікований на платформі *GitHub* та доступний за посиланням: https://github.com/EugeneSel/EUMT.
Ключові слова: метод; система нейронного машинного перекладу; системна інженерія; метод Еріксона–Пенкера.

Вступ

Поряд із надзвичайно широким застосуванням систем машинного перекладу (СМП) на ринку існує не так багато компаній-розробників, продукти яких мають попит. Це, зокрема, безоплатні та комерційні продукти, як-от: "*GoogleTranslate"*, *"DeepLTranslator"*, *"ModernMT"*, *"Apertium"* тощо. Практична потреба у просуванні на ринок якісних СМП потребує від розробників більш упорядкованих і систематизованих методів їх розроблення. Однак у літературі практично не висвітлюють хоч якісь загальноприйняті чи корпоративні стандарти й методи розроблення таких СМП. Водночас для більш ефективного та продуктивного розроблення якісних СМП потрібні науково обґрунтовані методи інженерії СМП, щоб якнайшвидше отримати якісний і конкурентоспроможний продукт.

У цій статті буде досліджено особливості розроблення систем нейронного машинного перекладу (СНМП), що поєднують найкращі властивості СМП на основі граматичних правил і СМП, реалізованих на статистичних алгоритмах.

Постановка задачі

Метою цієї статті є застосування бізнес-профілю Еріксона–Пенкера [1] для розроблення та формалізації методу системної інженерії СНМП.

Методологія системної інженерії та бізнес-профіль Еріксона–Пенкера

Основна ідея методу системної інженерії СНМП полягає у застосуванні методології системної інженерії та бізнес-профілю Еріксона–Пенкера для формалізації упорядкованого способу розроблення СНМП.

Методологія системної інженерії базується на трьох основних категоріях [2]:

1. Категорія “Система” як множина сутностей і відношень між ними, що в межах прийнятих припущень й обмежень показує, власне, систему.

2. Категорія “Життєвий цикл”, що передбачає представлення генезису системи від народження та до утилізації самої системи [2].

3. Категорія “Зацікавлені сторони”, що передбачає формування вичерпної множини умов і вимог, які зацікавлені сторони висувають до системи.

Проаналізуємо можливість застосування технік системної інженерії, яку також називають системною методологією XXI століття (за Дереком Хітчінсом [3]). Однією з найпоширеніших моделей представлення діяльності є бізнес-профіль Еріксона–Пенкера [1], в контексті якого автори сформулювали чотири основні сутності формального представлення діяльності будь-якої бізнес-системи:

– *цілі* (уособлюють мету діяльності системи та сформульовані як правило. Цілі можуть бути розбиті на підцілі та досягнені завдяки реалізації процесів);

– *процеси* (основні дії, що складають діяльність системи та призначені для досягнення мети відповідно до встановлених бізнес-правил. Процеси зазвичай підпорядковуються правилам, можуть змінювати стан вхідних ресурсів, а також продукувати нові ресурси – ресурси виходу системи згідно з умовами та вимогами, встановленими зацікавленими особами);

– *ресурси* (фізичні, абстрактні чи інформаційні об’єкти, які система споживає, використовує, обробляє та продукує впродовж всієї своєї діяльності для досягнення мети);

– *правила* (певні формалізовані обмеження, рамки, умови та вимоги тощо, що накладаються на процеси, а також описують характер зв’язків між ресурсами).

Основні діаграми, необхідні для формального графічного представлення та моделювання систем на основі бізнес-профілю Еріксона–Пенкера, поділяють на два основні типи:

– діаграми структурного представлення, а саме:

1. *Діаграму класів* (ієрархічний/логічний показ зв’язків і залежностей, наявних між класами сутностей, що складають систему. Під “класом сутностей” системи маємо на увазі ту чи іншу множину однієї з чотирьох сутностей, описаних вище);

2. *Діаграму компонентів* (схема поділу системи на частини – *компоненти*, що їх формують за певною, найчастіше функціональною, ознакою. Компоненти можуть реалізовувати один чи більше процесів і взаємодіяти з одним чи більше ресурсами. На діаграмі компонентів також вказують відношення між компонентами, реалізованими як формалізовані правила-*інтерфейси* для забезпечення взаємодії компонентів у системі);

– діаграми динамічного представлення, а саме:

1. *Діаграму діяльності* (зображує поетапний перебіг процесів системи, що пов’язані один з одним через вхідні та вихідні ресурсів; на діаграмі діяльності обов’язково зазначають її початок і кінець);

2. *Діаграму процесів із “водними доріжками”* (поєднує представлення процесів діяльності з компонентами системи, що реалізують ту чи іншу діяльність. Тобто будь-який процес має перебувати в межах свого батьківського компонента (відповідної вертикальної смуги, водної доріжки). Міжкомпонентні інтерфейси також можуть бути включені до цієї діаграми).

Така впорядкована множина формалізованих (зокрема, в нотації UML) сутностей і представлень системи на основі бізнес-профілю Еріксона–Пенкера є повною моделлю діяльності бізнес-системи, що враховує вимоги всіх зацікавлених осіб.

Метод системної інженерії СНМП

Під терміном “метод” тут і надалі у статті ми розуміємо “систематизований спосіб досягнення теоретичного чи практичного результату, розв’язання проблем чи одержання нової інформації на основі певних регулятивних принципів та дії, усвідомлення специфіки досліджуваної предметної галузі та законів функціонування її об’єктів” [4].

Метод розроблення СНМП (надалі – метод) на основі застосування технік системної інженерії складається з трьох основних етапів.

Формалізація структурного представлення СНМП. На першому етапі структуру СНМП моделюють як бізнес-профіль Еріксона–Пенкера [1, 5] (рис. 1): визначають її проблему, мету, ресурси, процеси, цілі, правила; проєктують структурне представлення в контексті, означеннях і моделях представлень області СМП з урахуванням інтересів усіх зацікавлених сторін.

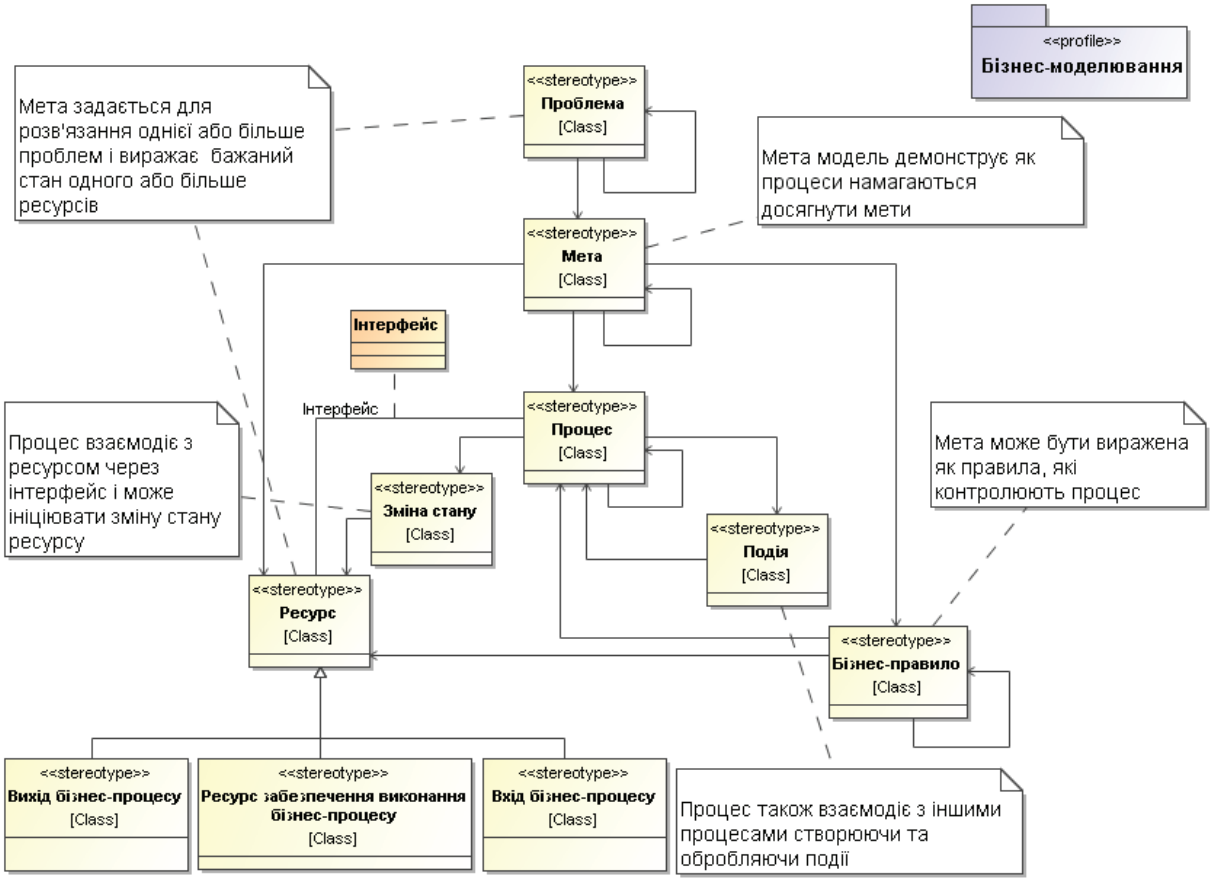

Рис. 1. Удосконалений бізнес-профіль Еріксона–Пенкера. Діаграма класів у нотації UML [5]

Визначимо зміст кожного з класів діаграми (див. рис. 1) у термінах постановки задачі інженерії СНМП, а саме класах:

1. *Проблема* (актуальне питання, що потребує відповідних рішень, основна мотивація розроблення СНМП, яка спонукає до формулювання конкретної мети. Проблема цієї роботи: *необхідність якісного англійсько-українського перекладу)*;

2. *Мета* (виражає глобальну ціль роботи, покликану розв'язати поставлену проблему. Мета цієї роботи: *розроблення конкурентоспроможного англійсько-українського перекладача)*;

3. *Процес* (множина процесів діяльності системи, внаслідок якої досягають мети, чітко визначена послідовність дій/підпроцесів, що призводить до виконання певного завдання. Процесами цієї системи є: *Завантаження текстових даних*, *Первинне оброблення текстових даних*, *Навчання методом машинного перекладу (ММП)*, *Машинний переклад (МП)*, *Функціонування вебзастосунка)*;

4. *Зміна стану* (можливі зміни певних ресурсів унаслідок роботи процесів. СНМП налічує три зміни станів: *Необроблені текстові дані* → *Оброблені текстові дані* (процес *Первинного оброблення текстових даних*), *Оброблені текстові дані* → *Перекладені текстові дані* (процес *МП*), *Ініціалізована ММП* → *Навчена ММП* (процес *Навчання ММП*));

5. *Ресурс* (будь-які сутності (матеріальні чи нематеріальні), що їх споживає та продукує розроблювана система. Детальнішу ієрархію ресурсів цієї системи наведено на рис. 2). Рис. 2 репрезентує компонентну модель системи англійсько-українського нейронного машинного перекладу (НМП), що відображає структуру та відношення між компонентами через реалізацію відповідних інтерфейсів.

Ресурси найнижчого рівня ієрархії, що беруть безпосередню участь у процесах, також поділяють за характером впливу на перебіг процесів на такі три класи:

– *Вихід бізнес-процесу* (ресурси, що їх продукує СНМП, кінцевий результат її функціонування. До них належать *Перекладені текстові дані*, *Навчена ММП)*;

– *Ресурс забезпечення виконання бізнес-процесу* (ресурси, що забезпечують виконання процесів, але не є кінцевим результатом роботи: *Оброблені текстові дані*, *Вебзастосунок*, *Модель текстового оброблення)*;

– *Вхід бізнес-процесу* (первинні ресурси входу початкових процесів, які ініціалізують цикл роботи системи: *Необроблені текстові дані*, *Ініціалізована ММП)*;

6. *Подія* (виникає через певні зовнішні фактори чи як результат взаємодії між процесами. Потенційними подіями цієї системи вважають зміну навчальних текстових даних, що впливає на *Навчання ММП*; появу нової найефективнішої версії *Навченої ММП* у процесі *Навчання ММП*, яка, як наслідок, замінить поточну версію, що бере участь у *МП*; введення текстових даних користувачем (запит користувача) в процесі *Функціонування вебзастосунка*, що спонукає систему до їх негайного перекладу);

7. *Бізнес-правило* ((з англ. – *Business Rule (BR)* формальні інструкції, що регулюють, обмежують, встановлюють контекст і рамки функціонування процесів. Приклад бізнес-правила СНМП: розмір введеного користувачем тексту для перекладу у вебзастосунку не може перевищувати 500 слів).

Формалізація динамічного представлення СНМП. На другому етапі визначають множину процесів, характерну саме для класу систем DataScience, згідно з визначеними О'Ніл і Шатт процесом DataScience [6] і міжнародним стандартом CRISP-DM в інтерпретації Фостера та Фосетта [7].

На цьому рівні організація діяльності може бути уточнена з урахуванням особливостей СНМП та, як наслідок, декомпонована на такі три підетапи:

1.1. Збір, аналіз й оброблення навчальних текстових даних, що їх використовуватимуть під час тренування нейронних мереж НМП відповідно до моделі процесу DataScience, запропонованої О'Ніл і Шатт [6], або стадії *Data under standing* і *Data processing* міжіндуструального стандарту процесів DataMining (*CRISP-DM — CR*oss *I*ndustry *S*tandard *P*rocess for *D*ata *M*ining) проаналізованого Фостером і Фосеттом [7].

1.2. Власне побудова (розроблення архітектури) та навчання нейронних мереж НМП – це аналог етапу *Machine Learning Algorithms Statistical Models* [6] чи стадії *Modeling* стандартного процесу для Data Mining [7].

1.3. Визначення метрик оцінювання ефективності як навчених моделей НМП, так і роботи системи загалом – аналог етапу *Report Findings* [6] чи стадії *Evaluation* стандартного процесу для Data Mining [7].

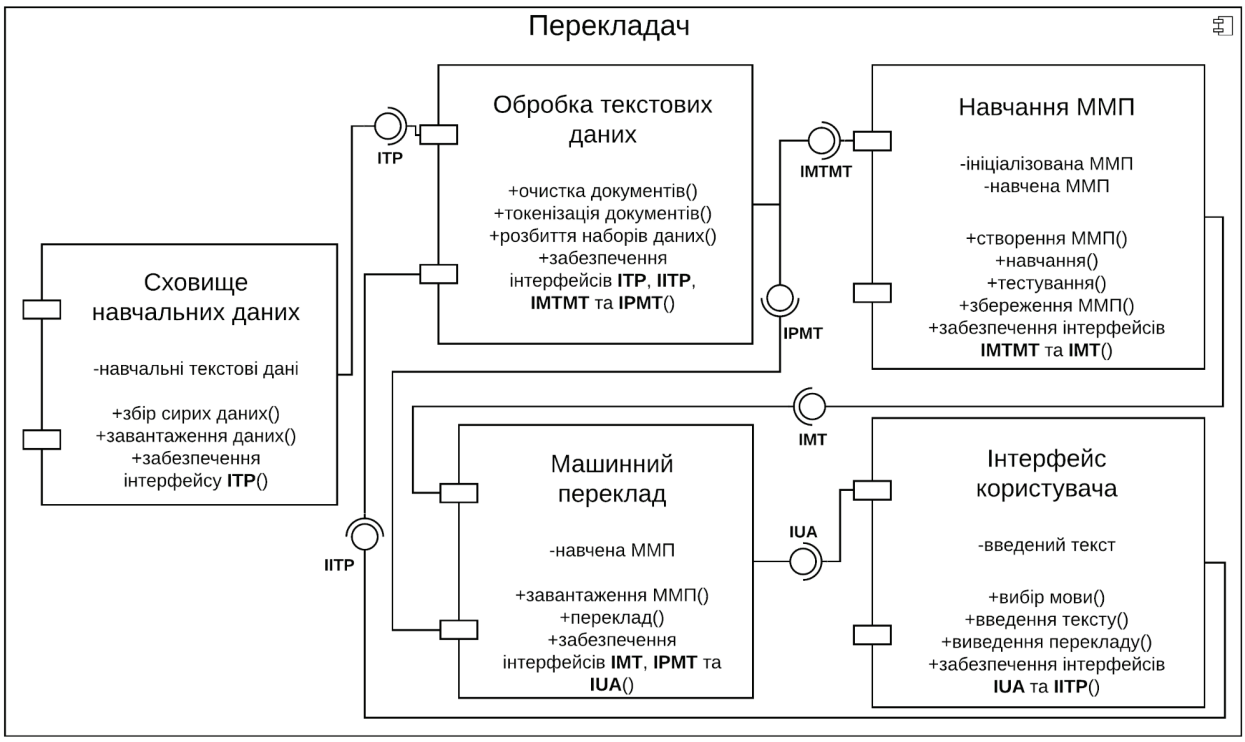

Рис. 2. Модель СНМП. Діаграма компонентів у нотації UML

Таким чином бізнес-профіль Еріксона–Пенкера є системою класів і відношень між ними, необхідних і достатніх для представлення та розроблення СНМП. Вичерпний перелік бізнес-правил (технічних умов) до бізнес-профілю регламентує функціональність інтерфейсу користувача конкретної СНМП, наприклад:

BR1 — Розмір введеного користувачем тексту для перекладу в Інтерфейсі користувача не може перевищувати 500 слів;

BR2 — Швидкість перекладу одного запиту користувача не повинна перевищувати 5 секунд;

BR3 — Графічний інтерфейс системи налічує 2 основні текстових поля: активне поле введення тексту мовою оригіналу, з яким взаємодіє користувач, а також поле виведення перекладеного тексту цільовою мовою;

BR4 — Функція зміни мов має бути імплементована в єдиній кнопці графічного інтерфейсу, враховуючи бінарний характер вибору.

Відповідна множина технологій Data Science є інструментами імплементації класів бізнес-профілю Еріксона—Пенкера.

Верифікація та валідація методу системної інженерії СНМП. На третьому етапі проводять верифікацію та валідацію розробленої СНМП задля перевірки дотримання всіх технічних умов і вимог зацікавлених сторін. Потрібно насамперед враховувати модель МП, яку застосовують для реалізації компонента ММП. Нею може бути один із типів архітектур нейронних мереж (цей перелік не є вичерпним):

– рекурентні нейронні мережі типу кодувальник/декодувальник із механізмом уваги [8];

– згорткові нейронні мережі з наскрізними зв'язками та згорткові нейронні мережі уваги [9];

– повнозв'язні нейронні мережі типу трансформер [10].

Далі можна провести аналіз функціональності, адекватності та продуктивності розробленої СМП на основі вибраної метрики порівняльного аналізу.

Тож, спираючись на означення поняття "метод" [4], можна сформулювати означення Методу системної інженерії СНМП: **це множина класів завдань, процесів, ресурсів, бізнес-правил і відношень між ними для продукування СНМП на основі методології системної інженерії, бізнес-профілю Еріксона–Пенкера та технологій Data Science.**

Імплементація методу системної інженерії СНМП

Далі покажемо застосування методу для реалізації СНМП.

Імплементація структурного представлення СНМП. Модель СНМП відповідно до представлень класів бізнес-профілю Еріксона–Пенкера (див. рис. 2) є системою класів і відношень між ними, необхідних і достатніх для розроблення СНМП.

Функціональність і призначення компонентів СНМП:

1. *Сховище навчальних даних.* Це локальне сховище, на якому розміщено зібрані корпуси текстів, використовувані для навчання моделі машинного навчання.

2. *Оброблення текстових даних.* Будь-які первинні текстові дані насамперед уніфікують. Набір текстових даних – так званий *корпус*, що складається з текстових *документів* – текстові рядки довільної довжини. Кожен документ первинного вхідного корпусу компонента оброблення текстових даних проходить такі етапи уніфікації:

2.1. *Очищення документа* від спеціальних символів, декапіталізація слів, аналіз абревіатур. Також можливе загальне очищення корпусу від невалідних документів.

2.2. *Токенізація документа* – розбиття текстового рядка на список атомарних текстових складових – *токенів*, що в цій імплементації є окремими словами.

2.3. *Розбиття навчальних текстових даних* на відповідні набори даних (тренувальний, валідаційний, тестовий).

Інтерфейс **IMTMT** – фактичне передання оброблених результівних навчальних текстових даних на вхід функції розбиття навчального текстового корпусу на тренувальний, валідаційний і тестовий набори даних компонента *Навчання ММП.*

Інтерфейс **IPMT** – фактичне передання оброблених результівних текстових даних, введених користувачем на вхід функції МП навченої моделі МП компонента МП.

3. *Навчання ММП.* Основний складник цього компонента – власне модель МП, якою може бути один із типів архітектур нейронних мереж (цей перелік не є вичерпним):

3.1. Рекурентні нейронні мережі типу кодувальник/декодувальник із механізмом уваги [8].

3.2. Згорткові нейронні мережі з наскрізними зв'язками та згорткові нейронні мережі уваги [9].

3.3. Повнозв'язні нейронні мережі типу трансформер [10].

Детальна архітектура нейронної мережі, а також усі відповідні гіперпараметри встановлюють емпірично під час безпосереднього етапу навчання з використанням валідаційного набору текстових даних.

У компоненті *Навчання ММП* передбачено функціонал:

– створення/ініціалізації ММП вищеописаної архітектури;

– її (пере)навчання з потенційною зміною архітектури та значень навчальних гіперпараметрів;

– фінальні якісне та кількісне тестування ефективності перекладу (проводять тільки після повного циклу валідації);

– збереження навченої ММП із застосуванням алгоритму контролю версій для подальшої підтримки та вдосконалення ММП.

Тут і надалі під версіями ММП необхідно розуміти ММП, навчені на різних версіях навчальних текстових даних, що з часом можуть бути змінені задля інтерпретації нововведених синтаксичних, семантичних, орфографічних правил тієї чи іншої мови, або ж доповнені з появою нових джерел даних перекладу. Такий контроль версій сприяє регулярному оновленню ММП згідно з останніми мовними стандартами.

4. *Машинний переклад.* Компонент безпосереднього НМП оброблених текстових даних, уведених користувачем (отримані через інтерфейс **IPMT**) через їх подачу на вхід завантаженої ММП (отриманої через інтерфейс **IMT**), що запущена в режимі передбачення (без зворотного проходження й оновлення ваг ММП). Завантаження останньої найефективнішої версії ММП входить до складу цього компонента як фактична імплементація інтерфейсу **IMT**. Він передає завантажену ММП на вхід процесу МП задля проведення операції НМП оброблених текстових даних, уведених користувачем.

Інтерфейс **IUA** – фактичне передання перекладених результівних текстових даних, уведених користувачем на вхід функції виведення перекладу компонента Інтерфейс користувача.

5. Інтерфейс користувача. Компонент, який уособлює елементарний графічний інтерфейс користувача, що йому надано три основні функції:

– обрати мову введення/перекладу;

– увести текст для перекладу: введені текстові дані передають інтерфейсом **IITP** на вхід функції очищення тексту компонента Оброблення текстових даних;

– переглянути результат перекладу (отриманий через інтерфейс **IUA**) – функція *виведення перекладу()* компонента Інтерфейс користувача.

Наведемо список усіх інтерфейсів СНМП:

– **ITP** (***Interface Text Processing***) – інтерфейс передання завантажених в оперативну пам'ять навчальних текстових даних на вхід процесу текстового оброблення;

– **IMTMT** (***Interface Machine Translation Model Training***) – інтерфейс передавання вихідних (із модуля текстового оброблення) опрацьованих (очищених, нормалізованих, токенізованих, нумеризованих) навчальних текстових документів на вхід процесу навчання ММП;

– **IPMT** (***Interface Processing-Machine Translation***) – інтерфейс передавання вихідних (із модуля текстового оброблення) опрацьованих (очищених, нормалізованих, токенізованих, нумеризованих) текстових даних, введених користувачем для перекладу, на вхід модуля безпосереднього НМП;

– **IMT** (***Interface Machine Translation***) – інтерфейс завантаження останньої найефективнішої версії збереженої навченої ММП для її подальшого використання в процесі НМП попередньо обробленого тексту, введеного користувачем;

– **IUA** (***Interface User-Application***) – інтерфейс передання результівного перекладу тексту користувача до модуля Інтерфейс користувача для його остаточного виведення;

– **IITP** (***Interface Input Text Processing***) – інтерфейс передавання тексту, введеного користувачем, до модуля первинного текстового оброблення.

Імплементація динамічного представлення системи. Діаграма діяльності (рис. 3) – це візуалізація основних циклів функціонування процесів і їхньої взаємодії першого рівня.

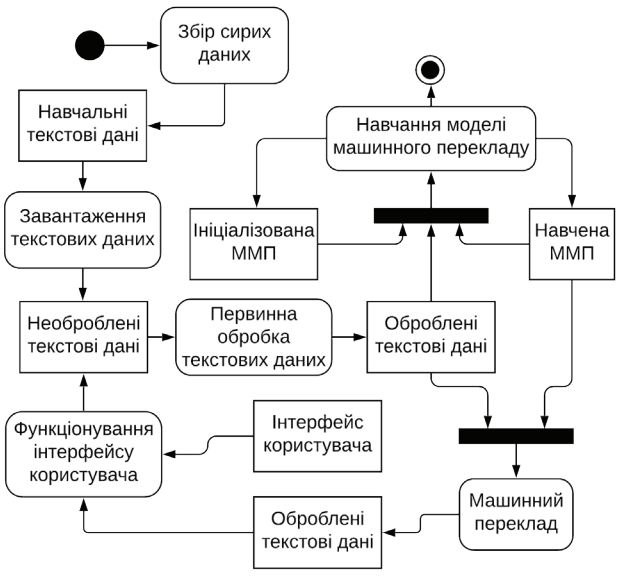

Рис. 3. Модель СНМП. Діаграма діяльності першого рівня в нотації UML

На рис. 4 показано діаграму процесів СНМП із виділенням “водних доріжок”, компонентів (див. рис. 2), отриману на основі вищенаведеної діаграми діяльності. Таке представлення дає змогу згрупувати основні процеси системи за відповідними компонентами. Діаграма процесів із водними доріжками є невіддільним складником моделі Еріксона–Пенкера, адже поєднує обидва типи представлень системи та формалізує її діяльність.

Імплементація компонентів Сховища навчальних й Оброблення текстових даних. Якість навчання будь-якої моделі машинного/глибинного навчання безпосередньо залежить від якості даних, залучених до тренування. Задля високої якості навчального текстового корпусу збирання та оброблення текстів повинні бути систематизовані. Детальний алгоритм збирання та оброблення навчальних текстових такий:

1. Збирання сирих даних:

– збирання/добирання сирих текстових даних із будь-яких відкритих ресурсів (використання відповідних *APIs*, завантаження наборів даних вручну);

– збереження текстових корпусів у табличному форматі: файли з розширенням .csv, .json, файли електронних таблиць *Excel* (.xls, .xlsx) тощо;

– елементарний опис зібраних корпусів: характеристики записів (*features*), кількість записів тощо.

2. Первинне оброблення текстових даних:

– підпроцес очищення документів:

- зведення даних до стандартних форматів (текстовий формат);
- видалення документів-дублікатів;
- усунення *NaN*, *Null* значень;
- очищення текстів від спеціальних символів, *HTML* тегів;
- декапіталізація слів, аналіз абревіатур;

– підпроцес токенізації документів: розбиття текстових рядків на список токенів, що в цій імплементації є окремими словами.

3. Розбиття набору даних на тренувальний, валідаційний і тестовий корпуси.

Імплементація компонента Навчання ММП. Процес Навчання ММП:

– побудова навчального словника токенів на основі тренувального корпусу;

– застосування моделі векторного представлення слів (наприклад, *sisg* [11]) з її попереднім тренуванням для показу токенів документів як векторів дійсних чисел (див. п. 3.1);

– створення ініціалізованої ММП початкової архітектури (див. п. 3.2) із використанням підручних програмних бібліотек / завантаження попередньо навченої ММП: останньої версії збереженої ММП або ж попередньо навченої ММП із зовнішніх ресурсів;

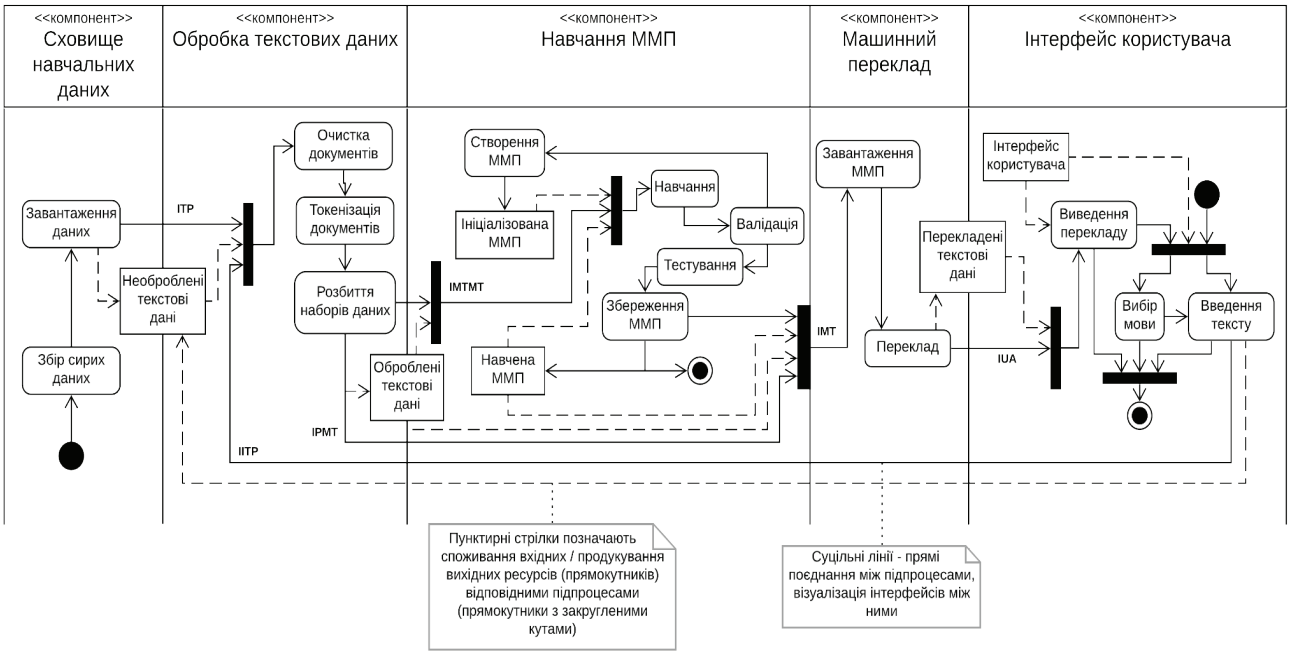

Рис. 4. Модель СНМП. Деталізована діаграма процесів на основі діаграми діяльності з водними доріжками другого рівня в нотації UML

– навчання ММП із застосуванням алгоритму оптимізації (наприклад, *Adam* [12]) на тренувальному корпусі $C^{train}C^{train}$ пар паралельних документів $\{X, Y\}\{X, Y\}$ оптимізацією цільової функції втрат (наприклад, класичної *перехресної ентропії* (*CE* (*C*ross-*E*ntropy)) [13]) послідовностей токенів *YY* (дійсний переклад *XX*) й $\hat{Y}\hat{Y}$;

– валідація для визначення оптимальної архітектури ініціалізованої ММП (тільки під час першої ітерації) та набору гіперпараметрів моделі на валідаційному корпусі $C^{val}C^{val}$ із підрахунком метрики *BLEU score* [14], описаної в п. 5;

– остаточне тестування на тестовому корпусі $C^{test}C^{test}$ навченої на $C^{train}C^{train}$ ММП з оптимальними архітектурою та набором гіпепараметрів із паралельним підрахунком і візуалізацією визначених метрик (функція втрат і *BLEU score* [14]);

– збереження поточної версії ММП, добір найефективнішої версії моделі з-поміж щойно навченої та попередньо збережених версій для безпосереднього МП запитів користувача.

Визначення критеріїв оцінювання продуктивності системи. Різні версії розроблюваної СНМП (відрізняються різними версіями застосовуваної ММП) порівнюватимуться з точки зору операбельності за часом оброблення (перекладу) запитів користувача. Потенційні кандидати на працездатну систему мають задовольняти *BR2* (див. п. 2.2).

Під час проведення підпроцесів валідації та тестування ММП процесу *Навчання ММП* як метрику оцінювання якості перекладу використовуватимуть *BLEU Score* [14]: отриманий переклад $\hat{Y}$ вхідної текстової послідовності X порівнюють із дійсним відповідним перекладом Y – документом, паралельним до X, наявним у застосовуваному текстовому корпусі. Кожна послідовність складається з токенів $X = \{x_i\}_{i=1}^{I}, Y = \{y_j\}_{j=1}^{J}, \hat{Y} = \{\hat{y}_k\}_{k=1}^{K}$ (де I, J та K – довжини X, Y і $\hat{Y}$ у токенах відповідно). З усіх токенів послідовності $\hat{Y}$ формують і підраховують усі можливі унікальні n-грами токенів довжиною від до N. Функція $Q_Y(g_n)$ – кількість появ n-грама g_n довжиною n у послідовності Y. Тоді для всіх послідовностей токенів довжини n вираховують таку величину [14]:

$$p_n = \frac{\sum_{g_n \in \hat{Y}} Q_{\hat{Y}}(g_n)}{\sum_{g_n \in \hat{Y}} Q_Y(g_n)} \forall n = \overline{1, N}. \tag{1}$$

Для всіх довжин n результати обчислень (2.2) комбінують в остаточну *BLEU Score* [14]:

$$BLEU = BP \cdot e^{\frac{1}{N}\sum_{n=1}^{N} \ln(p_n)}, \tag{2}$$

де BP – *Brevity Penalty* (дослівно – штраф стислості) [14]:

$$BP = \begin{cases} 1, & \text{якщо } K > J, \\ 1 - e^{1\frac{J}{K}}, & \text{інакше.} \end{cases} \tag{3}$$

З формул (1–3) очевидним є факт необхідності максимізації *BLEU Score*. Тож найкращу версію перекладача визначатимуть на основі найвищого значення саме цієї метрики.

Верифікація та валідація методу системної інженерії СНМП. На третьому етапі, завдяки запропонованому методу системної інженерії для СНМП, менш ніж за три місяці розроблено першу версію системи двоспрямованого англійсько-українського НМП на основі моделі *sisg* [11] векторного показу слів й архітектури нейронних мереж типу трансформер [10]. Різні версії трансформерів були навчені на корпусах паралельних англійських перекладів ресурсу *OPUS* [15]. Отримана система *EUTM* (*E*nglish-*U*krainian *M*achine *T*ranslator), яка за результатами порівняльного аналізу метрики BLEU Score, поданому в таблиці, щонайменше не поступається за якістю англійсько-українського перекладу загальнодоступному перекладачеві "*Google Translate"* [16].

Жирним шрифтом виділено найкращий показник для двох порівнюваних трансформерів (табл. 1).

***Таблиця* 1.** Порівняння якості перекладу навчених трансформерів із *"Google Translate"*

Версія трансформера (за використаними наборами даних)	**Кількість тренувальних епох**	***BLEU Score* англ-укр трансформера**	***BLEU Score* англ-укр *GT***	***BLEU Score* укр-англ трансформера**	***BLEU Score* укр-англ *GT***
Корпуси *WikiMatrix* [17] і *XLEnt* [18]	503 (4072825 унікальних навчальних документів)	**15.197**	7.772	7.452	**24.165**
Корпус *QED* [19]	5312 (197306 унікальних навчальних документів)	14.122	**14.329**	**20.932**	19.735
Корпус *Tatoeba* [15]	3055 (149038 унікальних навчальних документів)	**23.444**	22.899	**34.037**	10.071

За результатами оцінювання продуктивності систем продуктивності систем було встановлено, що розроблена система *EUMT*, розгорнута на віддаленому вебсервері безплатної платформи *MyBinder*, здатна обробляти запити користувача в межах однієї секунди, повністю задовольняючи встановлене бізнес-правило *BR2*.

Повний код розробленої версії системи EUMT опублікований на платформі *GitHub* та доступний за посиланням посиланням: https://github.com/EugeneSel/EUMT. Цей репозиторій містить також пряме посилання на вебзастосунок із можливістю його онлайн-тестування.

Висновки

1. Запропоновано метод системної інженерії СНМП, що базується на модифікованому бізнес-профілі Еріксона–Пенкера [1, 5] представлення системи на метарівні, а також міжнародних стандартів процесів DataScience [6] та DataMining [7], що є основою для алгоритмізації розроблення специфічних для НМП компонентів системи. Досліджено ефективність застосування методу на прикладі розроблення системи двоспрямованого англійсько-українського НМП EUMT і встановлено, що система EUMT щонайменше не поступається за якістю англійсько-українського перекладу популярному перекладачеві "*GoogleTranslate"*.

2. Запропонований метод системної інженерії СНМП спрямовано на розроблення вузькоспеціалізованих СНМП, призначених для первинного перекладу юридичних, медичних, бізнесових й інших важливих текстів, написаних семантично складними мовами, до яких належить й українська мова.

3. Формалізовано розроблення СНМП, що суттєво прискорює й упорядковує імплементацію СНМП і зменшує витрати на її створення.

4. Перспективи подальших досліджень спрямовані на застосування методу системної інженерії СНМП для реалізації СНМП на основі інших математичних моделей МП, формування метрик оцінювання продуктивності та науково обґрунтованих методів верифікації та валідації методу.

References

[1] H.-E. Eriksson and M. Penker, *Business modeling with UML*. New York: John Wiley & Sons, 2000, 459 p.

[2] A. Kossiakoff *et al.*, *Systems Engineering Principles and Practice*, V.K. Batovrin, Ed. Moscow, Russia: DMK Press, 2014, 624 p.

[3] D.K. Hitchins, *Systems Engineering: A 21st Century Systems Methodology*. Wiley, 2007, 528 p.

[4] S. Krymskyi, "Metod," in *Filosofskyi Entsyklopedychnyi Slovnyk*, V.I. Shynkaruk, Ed. Kyiv, Ukraine: Abrys, 2002, 742 p.

[5] P.P. Maslianko and O.S. Maystrenko, "The system engineering of organizational system informatization projects," *KPI Sci. News*, no. 6, pp. 34–42, 2008.

[6] C. O'Neil and R. Schutt, *Doing data science: Straight talk from the frontline.* O'Reilly Media, Inc., 2013, 406 p.

[7] F. Provost and T. Fawcett, *Data science for business: What you need to know about data mining and data-analytic thinking.* O'Reilly Media, Inc., 2013.

[8] D. Bahdanau *et al.*, "Neural machine translation by jointly learning to align and translate," in *3rd International Conference on Learning Representations*, San Diego, United States, 2014.

[9] J. Gehring *et al.* (2016). *A convolutional encoder model for neural machine translation* [Online]. Available: https://arxiv.org/pdf/1611.02344.pdf

[10] A. Vaswani *et al.*, "Attention is all you need," in *31st Conference on Neural Information Processing Systems*, Long Beach, CA, USA, 2017.

[11] P. Bojanowski *et al.*, "Enriching word vectors with subword information," *Trans. Assoc. Computation. Ling.*, vol. 5, pp. 135–146, 2017.

[12] D.P. Kingma and J.L. Ba. (2014). *Adam: A method for stochastic optimization* [Online]. Available: https://arxiv.org/pdf/1412.6980.pdf?source=post_page

[13] C. Szegedy *et al.* (2016). *Rethinking the inception architecture for computer vision* [Online]. Available: https://arxiv.org/pdf/1512.00567.pdf

[14] K. Papineni *et al.* "Bleu: A method for automatic evaluation of machine translation," in *Proc. 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, 2002, pp. 311–318.

[15] J. Tiedemann. (2012). *Parallel data, tools and interfaces in OPUS* [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf

[16] M. Johnson *et al.* (2017). *Google's multilingual neural machine translation system: Enabling zero-shot translation* [Online]. Available: https://arxiv.org/pdf/1611.04558.pdf

[17] S. Holger *et al.* (2019). *WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia* [Online]. Available: https://arxiv.org/pdf/1907.05791.pdf

[18] A. El-Kishky *et al.* (2021). *XLEnt: Mining Cross-lingual Entities with Lexical-Semantic-Phonetic Word Alignment* [Online]. Available: http://data.statmt.org/xlent/elkishky_XLEnt.pdf

[19] A. Abdelali *et al.*, "The AMARA Corpus: Building parallel language resources for the educational domain," in *Proc. 9th International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, 2014, p. 1856−1862.

П.П. Маслянко, Е.П. Сельский

МЕТОД СИСТЕМНОЙ ИНЖЕНЕРИИ СИСТЕМ НЕЙРОННОГО МАШИННОГО ПЕРЕВОДА

Проблематика. На рынке существует не так много компаний-разработчиков систем машинного перевода (СМП), продукты которых пользуются спросом. Это, например, "*GoogleTranslate"*, *"DeepLTranslator"*, *"ModernMT"*, *"Apertium"*, *"Trident"* и другие. Существует потребность в упорядоченных и систематизированных методах разработки СМП, а также нужны научно обоснованные методы инженерии систем нейронного машинного перевода (СНМП), чтобы как можно быстрее получить качественный и конкурентоспособный продукт.

Цель исследования. Применить бизнес-профиль Эриксона–Пенкера для разработки и формализации метода системной инженерии СНМП.

Методика реализации. Методология системной инженерии и бизнес-профиль Эриксона–Пенкера для формализации упорядоченного способа разработки СНМП.

Результаты исследования. Метод разработки СНМП на основе применения техник системной инженерии состоит из трех основных этапов.

На первом этапе структуру СНМП моделируют в виде бизнес-профиля Эриксона–Пенкера; на втором – определяют множество процессов, характерное для класса систем Data Science и международного стандарта CRISP-DM, а на третьем проводят верификацию и валидацию разработанной СНМП.

Выводы. Предложен метод системной инженерии СНМП, основанный на модифицированном бизнес-профиле Эриксона–Пенкера представления системы на метауровне, а также международных стандартов процессов Data Science и Data Mining. Исследована эффективность применения метода на примере разработки системы двунаправленного англо-украинского НМП EUMT (*E*nglish-*U*krainian *M*achine *T*ranslator) и установлено, что система EUMT по меньшей мере не уступает по качеству англо-украинского перевода популярному переводчику *"GoogleTranslate"*.

Полный код версии системы EUMT опубликован на платформе *GitHub* и доступен по ссылке: https://github.com/EugeneSel/EUMT.

Ключевые слова: метод; система нейронного машинного перевода; системная инженерия; метод Эриксона–Пенкера.

P.P. Maslianko, E.P. Sielskyi

METHOD OF SYSTEM ENGINEERING OF NEURAL MACHINE TRANSLATION SYSTEMS

Background. There are not many machine translation companies on the market whose products are in demand. These are, for example, free and commercial products such as "*GoogleTranslate"*, *"DeepLTranslator"*, *"ModernMT"*, *"Apertium"*, *"Trident"*, to name a few. To implement a more efficient and productive process for developing high-quality neural machine translation systems (NMTS), appropriate scientifically based methods of NMTS engineering are needed in order to get a high-quality and competitive product as quickly as possible.

Objective. The purpose of this article is to apply the Eriksson-Penker business profile to the development and formalization of a method for system engineering of NMTS.

Methods. The idea behind the neural machine translation system engineering method is to apply the Eriksson-Penker system engineering methodology and business profile to formalize an ordered way to develop NMT systems.

Results. The method of developing NMT systems based on the use of system engineering techniques consists of three main stages. At the first stage, the structure of the NMT system is modelled in the form of an Eriksson-Penker business profile.

At the second stage, a set of processes is determined that is specific to the class of Data Science systems, and the international CRISP-DM standard. At the third stage, verification and validation of the developed NMTS is carried out.

Conclusions. The article proposes a method of system engineering of NMTS based on the modified Erickson-Penker business profile representation of the system at the meta-level, as well as international process standards of Data Science and Data Mining. The effectiveness of using this method was studied on the example of developing a bidirectional English-Ukrainian NMTS EUMT (*E*nglish-*U*krainian *M*achine *T*ranslator) and it was found that the EUMT system is at least as good as the quality of English-Ukrainian translation of the popular Google Translate translator.

The full version code of the EUMT system is published on the *GitHub* platform and is available at: https://github.com/EugeneSel/EUMT.

Keywords: method; neural machine translation system; system engineering; Eriksson-Penker method.

Рекомендована Радою
факультету прикладної математики
КПІ ім. Ігоря Сікорського

Надійшла до редакції
10 квітня 2021 року

Прийнята до публікації
14 червня 2021 року