УДК 811.161.2

З.М. Палій, А.Б. Романюк
Національний університет “Львівська політехніка”,
кафедра систем автоматизованого проектування

ДОСЛІДЖЕННЯ АЛГОРИТМІВ АВТОМАТИЧНОГО ВИРІВНЮВАННЯ ТЕКСТІВ ДЛЯ АНГЛО-УКРАЇНСЬКОЇ МОВНОЇ ПАРИ

Описано проблему вибору програми автоматичного вирівнювання та наведені результати вирівнювання паралельних текстів при створенні паралельного англо-українського корпусу художньої літератури. Розглянуто процедуру попереднього оброблення вхідних текстів.

Ключові слова – лексикографія, алгоритм вирівнювання, ефективність

The research is mostly based on the search for a reliable alignment algorithm and its program implementation. As all the programs used within the scope of the research are freely-available and required some modifications to ensure their reliable performance, a number of supporting programs have been developed.

Keywords – lexicography, alignment algorithm, efficiency

Вступ

Паралельні корпуси текстів становлять важливе джерело емпіричних даних для проведення лінгвістичних досліджень, зокрема в сферах лексикографії та перекладознавства, для автоматичного оброблення природної мови, а також можуть використовуватися для ґрунтовного вивчення іноземних мов. Створення такого корпусу для англо-української мовної пари передбачає декілька етапів: збір двомовних матеріалів для включення в корпус, вибірка текстів однакового обсягу з різних джерел для забезпечення репрезентативності корпусу, вирівнювання текстів на одному з обраних рівнів, представлення текстів в єдиному обраному форматі, доступ до корпусу за допомогою корпусного менеджера. Це дослідження присвячене розгляду одного з найважливіших етапів створення паралельного корпусу тестів – автоматичному вирівнюванню текстів.

Вирівнювання паралельних текстів

На перший погляд може здатись, що переклад тексту означає збереження його внутрішньої структури, але практичне порівняння двох частин паралельного тексту показує, що структура тексту перекладу може бути не лише відмінною, але й зовсім іншою.

На рис. 1 наведено приклад складного вирівнювання за реченнями. Можна побачити, що: 9-те та 11-те речення тексту оригіналу перекладено одним 9-м реченням, 10-те речення пропущене у тексті перекладу, а 12-те речення оригіналу розбите перекладачем на три окремі речення (10, 11, 12).

У цьому випадку, якщо припустити, що порядкові номери речень тексту оригіналу відповідають порядковим номерам речень тексту перекладу, – 33% вирівнювань будуть помилковими, тобто кожне третє речення буде неправильно вирівняним, а отже, паралельний корпус, що міститиме таке вирівнювання не матиме ані теоретичної, ні практичної цінності. Для того, щоб забезпечити таку цінність, паралельні тексти всередині корпусу повинні бути вирівняні чи то людиною, чи то машиною. Людині, котра вирівнюватиме тексти зі здебільшого 100% точністю, знадобиться близько 5–7 хв для вирівнювання однієї сторінки паралельного тексту, а машині лише 0,07 с, хоча точність автоматичних вирівнювачів залежить від багатьох факторів, що розглядаються в межах цієї статті.

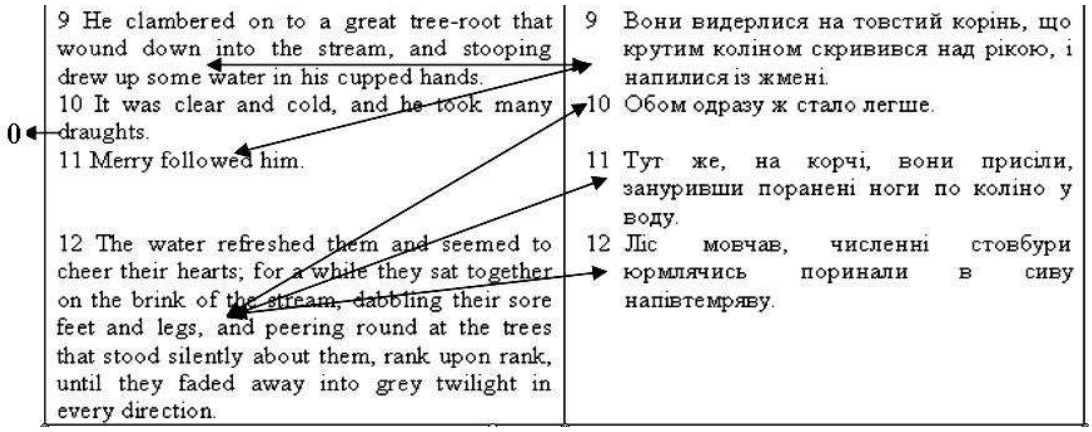

Рис. 1. Приклад вирівнювання речень паралельного тексту

До проблеми автоматичного вирівнювання вперше звернулися в середині 80-х років (дослідження проводилось в Xerox PARC, IBM, AT&T Bell Laboratories, ISSCO), а в середині 90-х з'явилися й численні публікації зі способами вирішення цієї проблеми (П. Браун (P. Brown) та співавтори у 1991, В. Ґейл (W. Gale) та К. Черч (K. Church) у 1991, В. Ґейл (W. Gale) та К. Черч (K. Church) у 1993, М. Сімард (M. Simard) у 1995, Д. Меламед (D. Melamed) у 1996, та ін.) [10]. Було розроблено різні підходи до автоматичного вирівнювання, які в основному різнились критеріями порівняння паралельних текстів та рівнями, на яких таке порівняння проводилось. Теоретично, вирівнювання паралельних текстів може проводитись на таких рівнях: цілий текст, глава, частина, абзац, речення, слово, символ.

На практиці здебільшого застосовується вирівнювання на рівні окремих речень та на рівні слів. Критеріями вирівнювання можна назвати інформацію, що береться за основу для порівняння двох (чи більше) текстів. Це може бути, наприклад, статистична інформація про довжини речень (тобто кількість слів або символів у реченні), лінгвістична інформація про слова-когнати (англ. cognate words), слова-ключі, якими можуть бути власні назви, числа або ж словникові входження (тобто слово вхідної мови та його переклад засобами вихідної мови в межах однієї словникової статті), частинномовна розмітка, а також інші засоби, властиві обраній мовній парі.

У межах цього дослідження увагу зосереджено на вирівнюванні паралельних текстів на рівні речень та проаналізовано роботу трьох основних (на думку авторів) алгоритмів для здійснення такого вирівнювання та програмних реалізацій обраних алгоритмів. Розглянуто:

1) алгоритм порівняння довжин, який опублікували В. Ґейл (W. Gale) та К. Черч (K. Church) у 1991 р.;

2) алгоритм пошуку слів-когнатів та словникових входжень, який опублікував Д. Меламед (D. Melamed) у 1996 р.;

3) алгоритм порівняння довжин та пошуку словникових входжень, який розробив для вирівнювання англо-угорського паралельного корпусу текстів Д. Варґа (D. Varga) та опублікований у 2005 р..

Наведені алгоритми використано у їхніх програмних реалізаціях:

Автоматичний вирівнювач Vanilla;

Автоматичний вирівнювач GMA;

Автоматичний вирівнювач hunalign.

Реферативний корпус текстів

Для оцінювання точності алгоритмів та роботи автоматичних вирівнювачів було створено реферативний корпус текстів, який складається з двох підкорпусів – Корпусу британської англійської мови (Корпус en-GB) та Корпусу американської англійської мови (Корпус en-US). До Корпусу британської англійської мови увійшла глава „Володаря Перснів" Дж.Р.Р. Толкієна та її переклад українською А. Немірової. Корпус американської англійської мови складається з п'яти частин твору „Постріли в ресторані Сірано" Р. Чандлера та його перекладу М. Пінчевського. Такий реферативний корпус був вирівняний вручну за реченнями для порівняння результатів вирівнювання з результатами роботи автоматичних вирівнювачів.

Корпус британської англійської вміщує 21 603 слова (431 речення) мовою оригіналу (англійською) та 14 652 слова (405 речень) мовою перекладу (українською). Результуючий паралельний текст нараховує 319 вирівняних пар речень. З усіх вирівнювань всередині Корпусу британської англійської, найчисленнішими виявилися вирівнювання типу 1-1 (тобто пара речень, де одному реченню оригіналу відповідає рівно одне речення перекладу). Такий тип вирівнювання зустрічався у 213 парах речень цього корпусу.

Корпус американської англійської нараховує 6 600 слів (529 речень) вхідною мовою та 6 266 слів (433 речення) вихідною мовою. Результуючий паралельний текст вміщує 417 вирівняних пар речень. Для цього корпусу вирівнювання типу 1-1 також були найчастотнішими та зустрічались у 326 парах речень. Усі типи вирівнювань, що зустрічались у підкорпусах та відповідна їм кількість пар речень наведена у табл.1.

Таблиця 1

Типи вирівнювань всередині реферативного корпусу

Тип вирівнювання	Кількість вирівнювань	
	Корпус британської англійської	Корпус американської англійської
1:1	213	326
1:0	16	2
1:2	14	9
2:1	32	55
2:2	9	6
1:3	2	2
3:1	6	8
3:3	2	-
n:1	10	-
1:n	2	-
n:m	18	1

Для найбільш точного встановлення відповідності між двома частинами паралельного тексту, речення були відокремлені одне від одного та пронумеровані. Для цього використовувалась програма SentenSplit, що була розроблена в межах цього дослідження.

Приклад кінцевого вигляду реферативного корпусу наведено на рис. 2.

```
11 12 | 10
 11 "What's the matter, Albert? Sick?"
 12 The boy worked a pale smile on his face.
 10 - Що з тобою, Елберте? Ти хворий? Хлопець через силу посміхнувся.

13 11
 13 "I'm workin' double shift. Corky's sick. He's got boils. I guess maybe I didn't eat enough."
 11 - Працюю другу зміну підряд. Вертун занедужав. Це в мене, мабуть, через те, що я давно не їв.

14 12
 14 The tall, brown-eyed man fished a crumpled five-spot out of his pocket, snapped it under the boy's nose.
 12 Високий карооκий чоловік витяг із кишені зіжмакану п'ятірку і сунув її хлопцеві під ніс.

15 16 | 13
 15 The boy's eyes bulged.
 16 He heaved upright.
 13 Той вирячив очі, відштовхнувся від стінки, випрямився.

17 14
 17 "Jeeze, Mister Carmady. I didn't mean-"
 14 - Та що ви, містере Молверн.

18 15
 18 "Skip it, Albert. What's a fin between pals? Eat some extra meals on me."
 15 - Бери, бери, Елберте. Що важливіше - дружба чи якась там п'ятірка? Вважай, що я частую тебе обідом.
```

Рис. 2. Кінцевий вигляд реферативного корпусу

Попереднє оброблення вхідних текстів

Для забезпечення оптимальної роботи автоматичних вирівнювачів, вхідні тексти були приведені до форми „одиниця вирівнювання (речення) на рядок”. Для розподілення тексту використовувалась вже зазначена програма SentenSplit.

У межах дослідження речення здебільшого розглядали як послідовність символів, що починається з великої літери та закінчується одним з кінцевих знаків пунктуації, як то „.”, „!”, „?” чи „...”. Проте на певному етапі виникла потреба розширити поняття „речення” як „одиниці вирівнювання”.

У [2] говориться: „Одна з найсуттєвіших складностей вирівнювання полягає в тому, що авторського розділення тексту на речення та абзаци не завжди дотримуються в тексті перекладу. Крім того, в різних мовах (а інколи й різних виданнях) прийняті різні способи графічного оформлення, що інколи ускладнює виявлення границь речення в автоматичному режимі. Порівняйте, наприклад, різні способи переходів від прямої мови персонажів до авторських ремарок.”

Наведемо приклад. В оригіналі наступні 3 речення поєднані в межах одного абзацу:

At last Merry halted. 'We can't go on like this,' he panted. 'I want some air.'

Слова автора (At last Merry halted та he panted) не виокремлюються від прямої мови персонажа. В українському перекладі цей абзац має такий вигляд:

Нарешті Меррі не витримав:

- Не можу більше! Давай хоча б дух переведемо.

Перший абзац містить лише слова автора, тоді як в другому наведено пряму мову персонажа.

Згідно вищенаведених прикладів можна припустити, що автоматичні вирівнювачі, що на першому етапі своєї роботи встановлюють відповідність між абзацами тексту (як наприклад вирівнювач Vanilla) не зможуть безпомилково вирівняти дані приклади. Тому вирішення цієї проблеми представлялось або у поєднанні українських речень в одне речення (тобто представлення комбінації речень за „одиницю вирівнювання на рядок”), або у розділенні англійського абзацу на два окремих Авторам видалось доцільнішим відокремити вступні слова автора від прямої мови персонажа. Так наприклад англійський варіант було розділено так:

At last Merry halted.

'We can't go on like this,' he panted. 'I want some air.'

щоб дати змогу автоматичному вирівнювачеві встановити наступні відповідності:

At last Merry halted. <=> Нарешті Меррі не витримав:

'We can't go on like this,' he panted. 'I want some air.' <=> – Не можу більше! Давай хоча б дух переведемо.

Для того, щоб речення всередині прямої мови персонажів не розділялись на окремі одиниці програмою SentenSplit, у програмі було додано кілька шаблонів прямої мови та описано такі речення за допомогою регулярних виразів мови програмування Python.

Наприклад для опису наступного шаблону, що зустрічається в тексті оригіналу 'I do not know. You might call it far, perhaps. But what does that matter?' використовувався такий регулярний вираз: **\'+.[^']+[.!?]+\'(?! \w).**

Для опису наступного шаблону з тексту перекладу: – Як вам сказати? Для вас, може, і далеко. А яке це має значення?, використовувався регулярний вираз виду **\-+ +[A-Z]+.+[.!?]**

Таким чином тексти були розділені на окремі речення. Хоча відсоток помилкових розділень є невеликим, все ж деякі речення доводилось корегувати вручну.

Хотілось би також звернути увагу на ще один аспект попередньої обробки текстів – транслітерацію. У той час як алгоритм В. Ґейла та К. Черча використовує лише статистичну інформацію про частини паралельного тексту, інші розглянуті алгоритми зосереджують свою увагу ще й на лінгвістичних критеріях, таких, як слова-когнати та слововходження з лексикону перекладу наведеного для використання вирівнювачем.

Слова-когнати (лат. cognati – рідні) – це слова, що мають спільне походження та схоже звучання в двох або більше самостійних мовах. [19] Розрізняють повні слова-когнати та часткові. Під повними розуміють пару слів двох споріднених мов, що походить від одного й того ж слова

прамови. Під частковими когнатами розуміють пару слів, що пишеться або читається однаково чи з незначними відмінностями. Так, наприклад, англійське слово resume та французьке слово resumé можна вважати когнатами. Часткові когнати також називають орфографічними когнатами. Англо-українська мовна пара – пара мов, що входять до різних мовних сімей, тому слова-когнати для цих мов є, як правило запозиченнями з однієї мови в іншу (з англійської в українську, як наприклад пари: scanner-сканер, microphone-мікрофон), або запозиченнями двох мов з іншої мови (наприклад, слова латинської мови, що зустрічаються і в англійській, і в українській мовах: forma-form-форма). Ідентифікація таких слів у паралельному тексті допомагає автоматичним вирівнювачам, що опираються на лінгвістичні параметри, встановити відповідність між двома частинами паралельного тексту. Важливість встановлення таких відповідностей зумовлена також відсутністю (чи принаймні невідомістю авторам) подібного електронного словника, що знаходився б у вільному користуванні і таким чином збільшив би точність та результативність даного дослідження.

Проте, хоча для людини, що знаходить слова-когнати в тексті, їхня спорідненість видається очевидною, комп'ютер не може встановити відповідність між ними, оскільки українська та англійська мова використовують різні абетки для написання слів. У [15] Д. Меламед пропонує використовувати так звані фонетичні когнати для мов, що використовують різні абетки, оскільки при запозиченні, як правило, слово зберігає свою фонетичну форму. Проте, за умови відсутності програмного забезпечення для порівняння фонетичних форм слів англійської та української мов, увагу було зосереджено на орфографічній формі. Для вирішення проблеми, пов'язаної з представленням слів засобами різних абеток, тексти української мови було попередньо транслітеровано згідно з [3].

Оцінка ефективності алгоритмів автоматичного вирівнювання

Для оцінки ефективності роботи автоматичних вирівнювачів використано такі показники, що застосовуються у сфері інформаційного пошуку:

Точність (Precision)

Повнота (Recall)

F-міра (F-measure)

У контексті вирівнювання речень ці показники визначаються такими формулами:

$$Точність = \frac{к-сть\ правильних\ вирівнювань}{к-сть\ запропонованих\ вирівнювань}$$

$$Повнота = \frac{к-сть\ правильних\ вирівнювань}{к-сть\ реферативних\ вирівнювань}$$

$$F-міра = 2 \times \frac{Точність \times Повнота}{Точність + Повнота}$$

У таблицях з результатами роботи вирівнювачів зазначені показники мають такі позначення: Точність – P; Повнота – R; F-міра – F.

Паралельні тексти з реферативного корпусу були автоматично вирівняні за програмами: Vanilla, GMA, hunalign.

Оцінювання було проведене для двох різних випадків: спочатку окремо оцінювались результати роботи вирівнювачів для всіх вирівняних речень, потім оцінювались результати для вирівнювань типу 1-1, оскільки ця група вирівнювань була найчастотнішою для обох підкорпусів реферативного корпусу. Крім того, в подальшому при застосуванні результатів цього дослідження для створення паралельного англо-українського корпусу текстів, саме вирівнювання типу 1-1, на думку авторів, є найбільш релевантним. Розглянемо такий приклад:

Якщо здійснити запит до реферативного корпусу за словом „cigarette", то при розгляді всіх типів вирівнювань можна отримати результат:

3 He got out of his LaSalle coupe and stood for a while by the side entrance to the Carondelet, the high collar of his blue suede ulster tickling his ears, his hands in his pockets and a limp **cigarette** sputtering between his lips. 4 Then he went in past the barbershop and the drugstore and the perfume shop

with its rows of delicately lighted bottles, ranged like the ensemble in the finale of a Broadway musical. 5 He rounded a gold-veined pillar and got into an elevator with a cushioned floor.

3 Вийшовши із свого двомісного «ласаля», він постояв трохи коло бічного входу до «Каронделета», тримаючи руки в кишенях просторого пальта із синьої замші, високий комір якого лоскотав йому вуха. 4 Коли намокла **сигарета** в його зубах зашипіла й погасла, Молверн увійшов до готелю, проминув перукарню, аптеку-закусочну, парфюмерну крамничку, на полицях якої вміло підсвічені пляшечки вишикувались, мов балетна трупа у фіналі бродвейського шоу, і, обійшовши колону із золотими прожилками, ступив у застелену килимом кабіну ліфта.

Коли розглядатимуться лише вирівнювання типу 1-1, отримаємо наступне:

48 He lit a **cigarette** and stood looking down at her.

44 Дивлячись на неї, Молверн запалив **сигарету**.

З першого прикладу важче встановити відповідності між словами у реченнях, а другий приклад є нагляднішим і простішим для сприйняття як людиною, так і машиною.

Ефективність роботи автоматичних вирівнювачів для всіх типів вирівнювань наведено у табл. 2.

Таблиця 2

Ефективність алгоритмів автоматичного вирівнювання для всіх типів вирівнювань

Вирівнювач	**Корпус британської англійської**			**Корпус американської англійської**		
	P	**R**	**F**	**P**	**R**	**F**
Vanilla	0.4	0.44	0.42	0.64	0.63	0.64
GMA	**0.53**	0.6	0.56	0.58	0.58	0.58
hunalign	**0.53**	**0.62**	**0.57**	**0.67**	**0.76**	**0.71**

З таблиці очевидно, що ефективність алгоритму автоматичного вирівнювача hunalign вища за ефективність інших проаналізованих вирівнювачів.

Результати ефективності роботи автоматичних вирівнювачів для вирівнювань типу 1-1 наведено у табл.3.

Таблиця 3

Ефективність алгоритмів автоматичного вирівнювання для вирівнювань типу 1-1

Вирівнювач	**Корпус британської англійської**			**Корпус американської англійської**		
	P	**R**	**F**	**P**	**R**	**F**
Vanilla	0.58	0.57	0.57	0.76	0.7	0.73
GMA	**0.73**	0.77	**0.76**	0.76	0.63	0.69
hunalign	0.63	**0.78**	0.7	**0.86**	**0.88**	**0.87**

Згідно з даними табл. 3, вирівнювачі GMA та hunalign показали найкращі результати для вирівнювань типу 1-1. Низькі результати для Корпусу британської англійської зумовлені великою кількістю пісень та віршів всередині прозового тексту. Більшість таких випадків призводять до помилкових вирівнювань, оскільки перекладачами не завжди (як, наприклад, у даному випадку) зберігається кількість рядків вірша чи пісні у тексті перекладу.

Згідно з вищенаведеними даними можна підсумувати, що використання автоматичного вирівнювача hunalign видається найдоцільнішим для автоматичного вирівнювання речень англо-української мовної пари.

Висновки

Метою дослідження було обрати найефективнішу програму автоматичного вирівнювання для англо-українських паралельних текстів. Для цього було обрано три алгоритми автоматичного вирівнювання та їхні програмні реалізації, та проаналізовано їхню ефективність для паралельних текстів художньої літератури англо-української мовної пари. Для оцінювання роботи автоматичних вирівнювачів було створено реферативний корпус текстів, що вміщує 736 пар речень вирівняних вручну, що використовувався як еталон для порівняння з вирівнюваннями здійсненими автоматичними вирівнювачами. Найкращих результатів досяг автоматичний вирівнювач hunalign з показниками Точність – 0,86; Повнота – 0,88; F-міра – 0,87. Хоча вдалі результати роботи були показані й вирівнювачем GMA з дещо нижчими показниками. Оскільки автоматичний вирівнювач hunalign не тільки показав високі результати роботи, але й виявився найпростішим у використанні, автори статті рекомендують його для здійснення вирівнювань речень англо-української мовної пари.

1. Волошин В.Г. Комп'ютерна лінгвістика: Навч. посібник. – Суми: ВТД „Університетська книга", 2004. – 382 с. 2. Добровольский Д. О., Кретов А. А., Шаров С. А. Корпус параллельных текстов: архитектура и возможности использования. – М.: Индрик, 2005. 3. Додаток до рішення № .9 української комісії з питань правничої термінології. Протокол № .2 від 19 квітня 1996 р. 4. Захаров В.П. Корпусная лингвистика: Учебно-метод. пособие. – СПб., 2005. 5. Толкієн Дж.Р.Р. Володар Перснів: Дві Вежі. – Харків: Фоліо, 2003. – 320 с. 6. Чандлер Р. Постріли в ресторані Сірано: Повісті. – К.: Всесвіт, 1992. – 208 с. 7. Широков В.А. та ін. Корпусна лінгвістика. – К.: Довіра. – 471 с. 8. Caseli H., Nunes V. Evaluation of Methods for Sentence and Lexical Alignment of Brazilian Portuguese and English Parallel Texts. – 13 p. 9. Danielsson P., Ridings D. Practical presentation of a "vanilla" aligner // TELRI Workshop in alignment and exploitation of texts. – 1997. 10. Davies A., Elder C. The handbook of applied linguistics. Blackwell Publishing Ltd. – 2004. 10. Gale W.A., Church K.W. A program for aligning sentences in bilingual corpora // Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL), Berkley – 1991. – P. 177–184. 11. Gale W.A., Church K.W.: A program for aligning sentences in bilingual corpora // Computational Linguistics 19, – 1993. – P. 75–10. 12. Melamed I.D. A geometric approach to mapping bitext correspondence // Proceedings of the Conference on Empirical Methods in Natural Language Processing, Philadelphia, Pennsylvania. – 1996 13. Melamed I.D. Automatic Evaluation and Uniform Filter Cascades for Inducing N-best Translation Lexicons // Proceedings of the Third Workshop on Very Large Corpora, Boston, MA – 1995 15. Melamed I.D.: A portable algorithm for mapping bitext correspondence // Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics. – 1997. – P. 305–312. 16. Simard M., Foster G., Isabelle P. Using Cognates to Align Sentences in Bilingual Corpora // Proceedings of TMI-92. – 1992 17. Singh A.K., Husain S. Comparison, selection and use of sentence alignment algorithms for new language pairs // Proceedings of the ACL – 2005 18. Varga D. et al. Parallel corpora for medium density languages – 2005. – P. 15–30. 19. Wikipedia entry on Cognate words – available from http://en.wikipedia.org/wiki/Cognate.