УДК 004.891 https://doi.org/10.17721/1812-5409.2023/1.10

Врублевський В.Н. [1], аспірант,
Марченко О. О.[1], д.ф.-м.н., проф.

Огляд підходів до розв'язання задач ідентифікації парафраз

[1]Київський національний університет імені Тараса Шевченка,83000, м. Київ, пр-т. Глушкова 4д,
e-mail: vitalii.vrublevskyi@gmail.com
rozenkrans17@gmail.com

V. N. Vrublevskyi, Post-graduate,
A. A. Marchenko[1], Dr. Sci., Prof.

Review of approaches for paraphrase identification

[1]Taras Shevchenko National University of Kyiv, 83000, Kyiv, Glushkova st., 4d,

e-mail: vitalii.vrublevskyi@gmail.com
rozenkrans17@gmail.com

Стаття присвячена огляду підходів до розв'язання задачі ідентифікації парафраз. Описується актуальність та використання даної задачі у таких задачах як виявлення плагіату, спрощення тексту та пошук інформації. Було розглянуто декілька класів вирішення даної задачі. Перший підхід заснований на ручних правилах - використовує вручну підібрані особливості базуючись на базових властивостях парафраз. Другий підхід заснований на лексичній подібності та різноманітних базах даних і онтології. Підходи, засновані на машинному навчанні також представлені у даній статті та описує архітектури, які можуть бути використані для ідентифікації парафраз. Останній підхід який розглянуто базується на глибокому навчанні та сучасних моделях трансформерів.

Також зроблено порівняльний аналіз корпусів даних, які використовують для розв'язання задачі ідентифікації.

Ключові слова: *обробка природної мови, ідентифікація парафраз, машинне навчання.*

The article is devoted to a review of approaches to solving the problem of identifying paraphrases. This problem's relevance and use in tasks such as plagiarism detection, text simplification, and information search are described. Several classes of solutions were considered. The first approach is based on manual rules - it uses manually selected features based on the fundamental properties of paraphrases. The second approach is based on lexical similarity and various databases and ontologies. Machine learning-based approaches are also presented in this paper and describe different architectures that can be used to identify paraphrases. The last approach considered is based on deep learning and modern models of transformers.

A comparative analysis of data corpora, which are used to solve the problem of identification paraphrases, is also done.

Keywords: natural language processing, paraphrase identification, machine learning.

Актуальність

Обробка природної мови стає все більш важливою в сферах технологій і бізнесу, оскільки задачі, які вона вирішує дозволяють обчислювальним системам ефективніше обробляти та розуміти людську мову.

Моделі природної мови є важливими для забезпечення ефективного спілкування між людьми, які розмовляють різними мовами. Системи машинного перекладу, розпізнавання мовлення - це лише декілька задач з цієї галузі, що допомагають покращувати нашу комунікацію.

Багато компаній використовують різні моделі природної мови для покращення обслуговування клієнтів: створення чат-ботів і віртуальних помічників. Вони розуміють запити клієнтів і

відповідають на них у режимі реального часу, наприклад Amazon Alexa [1]

Завдяки обробці природної мови ми можемо користуватися пошуковими системами і системами рекомендацій, що допомагають знаходити найбільш відповідні результати на наші запити.

Розпізнавання парафразів - важлива задача обробки природної мови, що має багато практичних застосувань.

- Виявлення плагіату. Ідентифікацію парафразів можна використовувати для виявлення випадків плагіату шляхом порівняння підозрілого тексту з великою базою корпусів, щоб побачити, чи містять вони перефразовані речення та фрази.
- Спрощення тексту. Парафрази використовують для заміни складних речень і фраз на простіші, які передають те саме значення. Це корисно як для обчислювальних систем, що неефективно працюють із важкими структурами тексту, так і для людей.
- Пошук інформації. Ідентифікацію парафразів використовують для підвищення якості пошукових систем. Вони ідентифікують документи, які містять не лише оригінальний пошуковий запит, а і його парафрази.

Тому розпізнавання парафразів є важливим завданням в обробці природної мови. Воно допомогає підвищити точність і ефективність програм, які покладаються на розуміння сенсу тексту.

Задача ідентифікації парафраз

Ідентифікація парафраз — це завдання визначення того, чи два тексти мають однакове або схоже значення, навіть якщо вони сформульовані по-різному.

Прикладом парафраз можуть слугувати такі речення:

- *Він розлютив мене, коли проявив свою невихованість за вечерею.*
- *Його неввічливість, плітки та загальна відсутність поваги за вечерею розлютили мене.*

У контексті обробки природної мови ідентифікація парафразу зазвичай розглядається як завдання бінарної класифікації, де вхідними даними є пара речень, а виходом є булеве значення, що вказує, чи є два речення парафразами, чи ні.

Вхідні речення можна представити різними способами залежно від конкретного завдання та наявних даних. Наприклад, їх можна представити у вигляді необробленого тексту або можуть бути попередньо оброблені та перетворені в більш структуроване представлення, таке як послідовність токенізованих слів або синтаксичне дерево речення.

Для розв'язання цієї задачі існують різноманітні алгоритми машинного навчання, включаючи традиційні моделі, такі як дерева рішень, метод опорних векторів (SVM)[2] логістична регресія, а також сучасні моделі глибокого навчання, такі як рекурентні нейронні мережі (RNN) [3] і моделі трансформатори на основі BERT [4] і GPT[5].

Ефективність моделей зазвичай оцінюється стандартними показниками, такими як точність, повнота, F1.

Загалом ідентифікація парафразу є складним завданням, яке вимагає глибокого розуміння семантики природної мови. Однак, за останні роки було досягнуто значного прогресу, і найсучасніші моделі досягають високого рівня точності на різних контрольних наборах даних.

Підходи до розв’язання задачі ідентифікації парафраз

Методи виявлення парафраз значно вдосконалилися протягом останніх років завдяки прогресу в обробці природної мови і машинному навчанні. Розглядаючи цю еволюцію можна виділити такі етапи, або класи методів.

- Підходи, засновані на ручних правилах: на початку досліджень багато систем базувалися на на правилах та евристиках створених вручну. Ці методи спиралися на синтаксичні та семантичні шаблони для ідентифікації парафраз і часто були обмежені їх залежністю від конкретних мовних особливостей.
- Використання лексичної подібності: дослідники почали використовувати лексичну подібність використовуючи такі системи як WordNet[6] і латентний семантичний аналіз (LSA)[7], для ідентифікації парафраз. Ці методи базувалися на ідеї, що слова зі схожими значеннями, як правило, зустрічаються в подібних контекстах.
- Контрольоване машинне навчання: оскільки методи машинного навчання стали більш популярними в обробці природної мови, дослідники почали розробляти контрольовані моделі навчання для виявлення парафраз. Ці моделі зазвичай використовують розмічені дані. Вони вивчають шаблони та особливості, які вказують на парафрази.

- Глибоке навчання: в останнє десятиліття глибоке навчання стало панівним підходом у машинному навчанні і застосовувалося для широкого кола завдань, включаючи виявлення парафраз. Моделі глибокого навчання, такі як нейронні мережі та трансформери, показали виняткову продуктивність у різних тестах.

Для того аби краще зрозуміти різні ідеї та алгоритми, ми розглянемо декілька підходів з кожного класу.

Досить складно знайти чи виділити підходи, які будуть використовувати лише один з вище наведених класів, тому при виборі алгоритмів ми класифікували їх за основною ідеєю.

Підходи, засновані на ручних правилах

Одним з прикладів підходу, який базується на ручних правилах слів є система iSTART [8]

Вона використовує різноманітні лексичні, синтаксичні та семантичні особливості для ідентифікації парафраз.

Для того визначити структуру парафразу, розглядається декілька підходів їх побудови:

1) Синоніми: заміна слів їх синонімами.
2) Тип речення: зміна типу речення з активного на пасивний або навпаки.
3) Словоформа/частина мови: зміна форми слова в іншу, наприклад змінити іменник на дієслово, прислівник або прикметник.
4) Розбиття речення: довге речення розбивається на маленькі.
5) Визначення: замінити слово на його визначення.
6) Структура речення: зміна структури речення.

В основі алгоритму лежить поняття графів концептів[9]. Для кожного речення будується свій граф і після цього відбувається порівняння графів двох кандидатів.

Процес порівняння полягає в тому, щоб знайти якомога більше збігів між трійками “поняття-відношення-поняття”.

Речення перетворюється на граф концептів за допомогою аналізатора Link Grammar [10].

Даний алгоритм було реалізовано і частина з описаних вище структур була використана для визначення парафраз, проте автори не використовували власний корпус даних для аналізу і не опублікували своїх результатів.

Підходи, засновані на лексичній подібності

Хорошим прикладом методу, де поєднана лексична та семантична подібність є система розроблена Раду Міхалчем та Андраш Чомай [11], яка була опубліковано у 2006 році.

У ньому поєднано два різних підходи до вимірювання семантичної подібності між двома частинами тексту: вимірювання на основі корпусу та вимірювання на основі знань. Властивості, засновані на корпусі, покладаються на статистичний аналіз великих текстових корпусів, тоді як властивості, засновані на знаннях, використовують зовнішні ресурси, такі як лексичні бази даних і онтології.

Автори презентують дві характеристики семантичної подібності на основі корпусу, засновані на дистрибутивній подібності та латентному семантичному аналізі.

Однією з них є характеристика *PMI-IR*. Вона базується на семантичній подібності слів у великому корпусі, була запропонована Терні [12].

$$PMI - IR(w_1, w_2) = \log_2 \frac{p(w_1 \,\&\, w_2)}{p(w_1) * p(w_2)}$$

Щоб слова вважалися близькими у корпусі використали вікно довжиною десять слів, як баланс між точністю та продуктивністю.

Також вони вводять шість характеристик, засновані на базі знань WordNet[6]. Наведемо декілька з них:

- Подібність за Лікокам і Ходоровом [13] базується на нормалізованій довжині шляху з урахуванням глибини загальної ієрархії

$$LH(w_1, w_2) = \log_2 \frac{len(w_1, w_2)}{2D},$$

де D – максимальна глибина ієрархії в базі знань.

- Подібність за Ву і Палмером [14] враховує глибину найменшого спільного предка концептів в ієрархії:

$$WP(w_1, w_2) = -\log_2 \frac{depth(LSO(w_1, w_2))}{depth(w_1) + depth(w_2)},$$

де LSO – найменший спільний предок концептів пари слів.

Результати експериментів показують, що властивості як на основі корпусу, так і на основі знань можуть бути ефективними для ідентифікації парафраз.

Автори оцінюють систему використовуючи Microsoft Research Paraphrase Corpus[15]. Результати показують, що даний метод досягає точності 70.3% та F1 81.3 %.

Підходи, засновані на машинному навчанні

Варто зазначити, що більшість підходів так чи інакше використовують різний інструментарій машинного навчання, в цьому розділі ми розглянемо метод, де це є основою алгоритму.

У своїй статті Нітін Маднані та Джоель Тетро [16] пропонують новий підхід до оцінки моделей ідентифікації парафразів. Вони перепрофільовують чинні метрики машинного перекладу. Автори стверджують, що теперішні метрики машинного перекладу, такі як BLEU, METEOR і TER, добре підходять для побудови моделей ідентифікації парафраз.

Розглянемо декілька таких метрик:

1. BLEU [17] одна з найбільш часто використовуваних метрик для оцінки машинного перекладу. Вона базується на кількості однакових n-грам різних довжин у двох реченнях.

$$BLEU(s_1, s_2) = BP(s_1, s_2)\exp\left[\sum_{n=1}^{N}\frac{1}{N}\log(p_n)\right],$$

$$BP(s_1, s_2) = \exp\left[\min[1-\frac{|s_1|}{|s_2|},1]\right],$$

$$p_n = \frac{\sum_{x\in NGram(s_1,n)} count(x, NGram(s_1,n)\cap NGram(s_2,n))}{\sum_{x\in NGram(s_1,n)} count(x, NGram(s_1,n))},$$

$$count(x,S) = |\{el \mid el\in S \,\&\, el = x\}|,$$

N – *максимальна довжина n – грами*

де BP – коефіцієнт стислості, призначений для штрафування речень, якщо одне коротше за інше.

2. NIST[18] є варіантом BLEU, яка використовує середнє арифметичне кількості однакових n-грам, а не середнє геометричне. Ця метрика також зважує кожен n-грам відповідно до його інформативності, використовуючи її частоту.
3. TER[19] визначається як кількість редагувань, необхідних для того, аби з одного речення зробити інше використовуючи операції вставки, видалення та заміни та зсуву.

Ансамбль декількох класифікаторів використовується для вирішення поставленої задачі. На верхньому рівні – простий мета класифікатор, який використовував результати класифікації декількох інших класифікаторів нижнього рівня: логістичну регресію, метод опорних векторів та класифікатор на основі алгоритму найближчого сусіда [20].

Автори оцінюють систему використовуючи Microsoft Research Paraphrase Corpus [15]. Результати показують, що метод досягає точності 77.4% та F1 84.1%.

Підходи, засновані на глибокому навчанні

Більшість сучасних підходів використовують переваги великих розмічених корпусів даних. Нейронні мережі й трансформери застосовують в якості алгоритмічного фундаменту.

Одним з яскравих представників таких алгоритмів є система Sentence-BERT [21]

У цій статті пропонується новий метод створення високоякісних представлень речень за допомогою сіамської мережі та BERT.

Представлення речення – це вектор фіксованої довжини, який використовується для того, аби описати семантичне значення речення у багатовимірному векторному просторі. Вони застосовуються в різних завданнях обробки природної мови, включаючи ідентифікацію парафразів, класифікацію тексту та пошук інформації.

Автори починають із зауваження, що наявні методи генерації вбудованих речень зазвичай використовують неконтрольовані підходи, такі як усереднення представлення слів та інші. Однак ці методи часто призводять до низькоякісних представлень через природну складність захоплення семантичного значення речення з простого набору слів.

Мережа приймає як вхідні дані два речення та виводить оцінку схожості між ними, яку можна використовувати для створення високоякісних векторів, що будуть описувати речення.

Автори спочатку навчають сіамську мережу BERT на великому наборі даних пар речень, позначених бінарними мітками, використовуючи функцію втрат. Функція втрати мінімізує відстань між парафразами та максимізує відстань між не-парафразами у просторі.

Після навчання Нітін Маднані та Джоель Тетро використовують мережу BERT для створення представлення речень для різноманітних завдань, включаючи ідентифікацію парафразів. Результати показують, що модель Sentence-BERT перевершує чинні найсучасніші методи на більшості контрольних наборів даних.

Корпуси даних

Таблиця 1

Корпуси даних

Назва корпусу	Рік створення	Джерела данних	Розмір
Microsoft Research Paraphrase Corpus [15]	2005	Статті новин, енциклопедій і веб-сторінки	5,801
PPDB [22]	2013	Статті новин, веб-сторінки і державні документи	220,000,000
Sentences Involving Compositional Knowledge [23]	2014	Статті про політику, науку та розваги.	10,000
Пари запитань Quora [24]	2017	Пари запитань з веб-ресурсу	404,289
ParaNMT-50M [25]	2017	Статті новин, веб-сторінки і державні документи	51,409,585

Існує досить багато різних корпусів, які зазвичай використовуються для задачі ідентифікації парафразу. Наведемо деякі з них:

1) Microsoft Research Paraphrase Corpus (MSRPC) [15] – це корпус пар речень, для яких вручну зазначили чи є вони парафразами. Долан і Брокетт створили його у Microsoft Research і він став одним із найпоширеніших наборів даних для даної задачі. Корпус містить 5801 пару речень, які взяті з різних джерел, включаючи статті новин, статті в енциклопедіях і веб-сторінки. Щоб забезпечити якість міток, для кожної пари речень використовували кілька анотаторів, а розбіжності вирішувалися більшістю голосів.
2) PPDB [22]- це масштабна база даних, з понад 220 мільйонів пар парафраз. Вона містить багато фразових та лексичних парафразів.
3) SICK (Sentences Involving Compositional Knowledge) [23] — це корпус для задач семантичної близькості. Він містить 10,000 пар речень політичні, наукові та розважальні теми. Пари були підібрані таким чином, щоб вони мали різний ступінь семантичної спорідненості, починаючи від дуже схожих до зовсім не пов’язаних. Семантичний зв’язок між парами речень оцінюється за безперервною шкалою від 1 (зовсім не пов’язані) до 5 (семантично еквівалентні).
4) Пари запитань Quora [24] - це набір даних пар запитань із веб-сайту Quora. Мета набору даних — визначити, чи є два запитання семантично еквівалентними. Питання охоплюють широкий спектр тем, зокрема науку, технології, політику та розваги. Він містить понад 400 000 пар запитань і є одним із найбільших корпусів, доступних для ідентифікації парафраз.
5) ParaNMT-50M [25]: це великий корпус паралельних речень із різних джерел. Він містить понад 50 мільйонів пар речень і є одним із найбільших доступних наборів даних. Пари речень у ParaNMT-50M узяті з різних джерел, у тому числі з новинних статей, вебсторінок і державних документів. Корпус згенерований автоматично за допомогою перекладу частини декількох чесько-англійських корпусів.

Вище ми навели корпуси, що найчастіше використовуються для вирішення цієї задачі. Вибір набору даних часто залежить від конкретного завдання та контексту, у якому використовуватиметься модель, а також від обчислювальних потужностей, які є в дослідників.

Список використаних джерел

1. Chatbots in Call Centers [Електронний ресурс] – Режим доступу до ресурсу:

https://aws.amazon.com/chatbots-in-call-centers/.
2. *Cortes C.* Support-vector networks / C. Cortes, V. Vapnik. // Mach Learn. – 1995. – №20. – С. 273–297. https://doi.org/10.1007/BF00994018
3. *Yin, W., Kann, K., Yu, M., Schutze, H.* Comparative Study of CNN and RNN for Natural Language Processing – Режим доступу до статті: https://arxiv.org/abs/1702.01923
4. *Devlin J., Chang M.-W., Lee K., Toutanova K.* BERT: Pre-training of deep bidirectional transformers for language understanding – Режим доступу до статті: https://arxiv.org/pdf/1810.04805.pdf
5. *Radford A., Narasimhan K.* Improving Language Understanding by Generative Pre-Training
6. Fellbaum C. WordNet: An Electronic Lexical Database / Christiane Fellbaum., 1998. – (The MIT Press). https://doi.org/10.7551/mitpress/7287.001.0001
7. *Landauer, T., Foltz, P., & Laham, D.* An Introduction to Latent Semantic Analysis – Режим доступу до статті: https://doi.org/10.1080/01638539809545028
8. *Boonthum, C.* iSTART: Paraphrase Recognition – Режим доступу до статті: https://aclanthology.org/P04-2006.pdf
9. Sowa, J. F. Conceptual graphs as a universal knowledge representation – Режим доступу до статті: https://core.ac.uk/download/pdf/82803127.pdf
10. *Sleator, D., & Temperley, D.* Parsing English with a Link Grammar. – Режим доступу до статті: https://www.researchgate.net/publication/220482445_Parsing_English_with_a_Link_Grammar
11. *Mihalcea, R., Corley, C., & Strapparava, C.* Corpus-based and Knowledge-based Measures of Text Semantic Similarity – Режим доступу до статті: https://cdn.aaai.org/AAAI/2006/AAAI06-123.pdf
12. *Turney, P. D.* Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL – Режим доступу до статті: https://arxiv.org/abs/cs/0212033
13. *Leacock, C., Chodorow, M., & Miller, G. A.* Using Corpus Statistics and WordNet Relations for Sense Identification – Режим доступу до статті: https://aclanthology.org/J98-1006
14. *Wu Z. and Palmer M.* Verb semantics and lexical selection // 32nd Annual Meeting of the Association for Computational Linguistics. – New Mexico State University, Las Cruces, New Mexico. – 1994. – С. 133 – 138.
15. *Dolan, B., Quirk, C., & Brockett, C.* Unsupervised Construction of Large Paraphrase Corpora: Exploiting Massively Parallel News Sources – Режим доступу до статті: https://aclanthology.org/C04-1051.pdf
16. *Madnani, N., Tetreault, J., & Chodorow, M.* Re-examining Machine Translation Metrics for Paraphrase Identification – Режим доступу до статті: https://aclanthology.org/N12-1019.pdf
17. *Papineni K., Roukos S., Ward T., Zhu W.* BLEU: a Method for Automatic Evaluation of Machine Translation – Режим доступу до статті: https://www.aclweb.org/anthology/P02-1040.pdf
18. *Doddington G.* Automatic Evaluation of Machine Translation Quality using N-gram Co-occurrence Statistics / Doddington. // Proceedings of the second international conference on Human Language Technology Research. – 2002. – С. 138–145.
19. S*nover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J*. A Study of Translation Edit Rate with Targeted Human Annotation – Режим доступу до статті: https://aclanthology.org/2006.amta-papers.25
20. *Aha, D. W., Kibler, D., & Albert, M. K.* Instance-based learning algorithms – Режим доступу до статті: https://www.researchgate.net/publication/220343419_Instance-Based_Learning_Algorithms
21. *Reimers, N., & Gurevych, I.* Sentence-BERT: Sentence Embeddings using Siamese BERT – Режим доступу до статті: https://aclanthology.org/D19-1410.pdf
22. *Ganitkevitch, J., Van Durme, B., & Callison-Burch, C.* PPDB: The Paraphrase Database – Режим доступу до статті: https://aclanthology.org/N13-1092.pdf
23. *Marelli, M., Bentivogli, L., Baroni, M., Bernardi, R., Menini, S., & Zamparelli, R.* SemEval-2014 Task 1: Evaluation of Compositional Distributional Semantic Models on Full Sentences through Semantic Relatedness and Textual Entailment – Режим

доступу до статті: https://aclanthology.org/S14-2001.pdf
24. Quora Duplicate Questions | Kaggle [Електронний ресурс] – Режим доступу: https://www.kaggle.com/aymenmouelhi/quora-duplicate-questions
25. *Wieting, J., & Gimpel, K.* ParaNMT-50M: Pushing the Limits of Paraphrastic Sentence Embeddings with Millions of Machine Translations – Режим доступу до статті: https://aclanthology.org/P18-1042.pdf

References

1. AMAZON WEB SERVICES, INC. (2019). Chatbots in Call Centers – Amazon Web Services (AWS). [online] Available at: https://aws.amazon.com/chatbots-in-call-centers/.
2. CORTES, C. and VAPNIK V. (1995). Support-vector networks. Machine learning, 20(3), pp.273–297.
3. YIN, W., KANN, K., YU, M., & SCHÜTZE, H. (2017). Comparative Study of CNN and RNN for Natural Language Processing.
4. DEVLIN, J., CHANG, M.-W., LEE, K., & TOUTANOVA, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv. https://doi.org/10.48550/ARXIV.1810.04805
5. RADFORD, A., & NARASIMHAN, K. (2018). Improving Language Understanding by Generative Pre-Training.
6. FELLBAUM, C. (1998). WordNet. An Electronic Lexical Database.
7. LANDAUER, T., FOLTZ, P., & LAHAM, D. (1998). An Introduction to Latent Semantic Analysis. Discourse Processes, 25, 259–284. https://doi.org/10.1080/01638539809545028
8. BOONTHUM, C. (2004). iSTART: Paraphrase Recognition. Proceedings of the ACL Student Research Workshop, 31–36. https://aclanthology.org/P04-2006
9. SOWA, J. F. (1992). Conceptual graphs as a universal knowledge representation. Computers & Mathematics with Applications, 23(2), 75–93.
10. SLEATOR, D., & TEMPERLEY, D. (1995). Parsing English with a Link Grammar. CoRR, abs/cmp-lg/9508004.
11. MIHALCEA, R., CORLEY, C., & STRAPPARAVA, C. (2006). Corpus-based and Knowledge-based Measures of Text Semantic Similarity. Proceedings of the National Conference on Artificial Intelligence, 1.
12. TURNEY, P. D. (2001). Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL. Proceedings of the 12th European Conference on Machine Learning, 491–502.
13. LEACOCK, C., CHODOROW, M., & MILLER, G. A. (1998). Using Corpus Statistics and WordNet Relations for Sense Identification. Computational Linguistics, 24(1), 147–165. https://aclanthology.org/J98-1006
14. WU, Z., & PALMER, M. (1994). Verb Semantics and Lexical Selection. 32nd Annual Meeting of the Association for Computational Linguistics, 133–138. https://doi.org/10.3115/981732.981751
15. DOLAN, B., QUIRK, C., & BROCKETT, C. (2004). Unsupervised Construction of Large Paraphrase Corpora: Exploiting Massively Parallel News Sources. COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics, 350–356. https://aclanthology.org/C04-1051
16. MADNANI, N., TETREAULT, J., & CHODOROW, M. (2012). Re-examining Machine Translation Metrics for Paraphrase Identification. Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 182–190. https://aclanthology.org/N12-1019
17. PAPINENI, K., ROUKOS, S., WARD, T., & ZHU, W.-J. (2002). Bleu: a Method for Automatic Evaluation of Machine Translation. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 311–318. https://doi.org/10.3115/1073083.1073135
18. DODDINGTON, G. R. (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics.
19. SNOVER, M., DORR, B., SCHWARTZ, R., MICCIULLA, L., & MAKHOUL, J. (2006). A Study of Translation Edit Rate with Targeted Human Annotation. Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers, 223–231. https://aclanthology.org/2006.amta-papers.25
20. AHA, D. W., KIBLER, D., & ALBERT, M. K. (1991). Instance-based learning algorithms. Machine Learning, 6(1), 37–66.

21. REIMERS, N., & GUREVYCH, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 3982–3992. https://doi.org/10.18653/v1/D19-1410
22. GANITKEVITCH, J., VAN DURME, B., & CALLISON-BURCH, C. (2013). PPDB: The Paraphrase Database. Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 758–764. https://aclanthology.org/N13-1092
23. MARELLI, M., BENTIVOGLI, L., BARONI, M., BERNARDI, R., MENINI, S., & ZAMPARELLI, R. (2014). SemEval-2014 Task 1: Evaluation of Compositional Distributional Semantic Models on Full Sentences through Semantic Relatedness and Textual Entailment. Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), 1–8. https://doi.org/10.3115/v1/S14-2001
24. KAGGLE. (2017) Quora Duplicate Questions [Online] – Available from: https://www.kaggle.com/aymenmouelhi/quora-duplicate-questions.
25. WIETING, J., & GIMPEL, K. (2018). ParaNMT-50M: Pushing the Limits of Paraphrastic Sentence Embeddings with Millions of Machine Translations. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 451–462. https://doi.org/10.18653/v1/P18-1042

Надійшла до редколегії 15.04.2023