УДК 004.021:81’32
DOI https://doi.org/10.32851/tnv-tech.2021.6.5

ДОСЛІДЖЕННЯ МЕТОДІВ ВЕКТОРИЗАЦІЇ ТЕКСТІВ У ЗАДАЧАХ ВАЛІДАЦІЇ ВІДПОВІДЕЙ, ПОДАНИХ ПРИРОДНОЮ МОВОЮ

Кузьма К.Т. *– кандидат технічних наук,*
старший викладач кафедри інформаційних технологій
Миколаївського національного університету імені В.О. Сухомлинського
ORCID ID: 0000-0002-0937-7299
Мельник О.В. *– кандидат технічних наук,*
завідувач кафедри економіки та інформаційних технологій
відокремленого структурного підрозділу закладу вищої освіти
«Відкритий міжнародний університет розвитку людини «Україна»
Миколаївського інституту розвитку людини
ORCID: 0000-0002-9778-4109

Інтелектуалізація процесу обробки природномовних текстів у задачах автоматизованого тестування зумовлює актуальність дослідження. Оскільки відповіді відкритого типу в системах тестування є природномовними текстами, то завдання їх обробки належить до прикладної задачі обробки текстів. Усі прикладні задачі обробки текстів, рішення яких відбувається з використанням машинного навчання, нейромереж, вимагають векторизації – перетворення тексту на цифрові послідовності. Метою статті є дослідження моделей, методів векторизації текстів у задачах обробки відповідей, поданих природною мовою.

На першому етапі досліджено базові прикладні задачі обробки текстів і наведено їх класифікацію. Обґрунтовано віднесення задачі перевірки природномовних відповідей у межах цього дослідження до задач класифікації текстів і семантичного аналізу.

На другому етапі проаналізовано базові моделі представлення тексту в цифровому вигляді: bag-of-words та дистрибутивну семантику. Обґрунтовано застосування моделі bag-of-words для задачі обробки відповідей відкритого типу, оскільки для визначення класу відповіді досить складу словника, який застосовується для кодування колекції правильних відповідей, і частоти слів, з якою вони застосовуються у відповідях «навчального» та «тестового» наборів даних. Зауважено, що вектором ознак у цій задачі є частоти появи токенів (символьні або словесні уні-, бі-, n-грами) словника, сформованого за навчальною вибіркою, у відповідях «навчального» та «тестового» наборів даних.

На третьому етапі досліджено підходи до обчислення вектора ознак: абсолютну частоту (TF), відносну частоту (TF-IDF), сумісну інформацію (PWI), визначено переваги та недоліки кожного з них.

На останньому етапі для векторизації текстів у задачах обробки відповідей, поданих природною мовою, запропоновано такі комбінації наборів ознак: модель bag-of-words та TF; модель bag-of-words та TF-IDF; словесні n-грами та TF-IDF; символьні n-грами та TF-IDF; модель bag-of-words та TF-PWI.

Запропоновані набори ознак та їх комбінації є засобами покращення моделі машинного навчання для задачі перевірки відповідей, поданих природною мовою. Подальші дослідження будуть спрямовані на розробку моделі машинного навчання цієї задачі та її експериментальне тестування із запропонованими наборами ознак для отримання ефективної математичної моделі.

Ключові слова: *відповідь, подана у текстовій формі, природномовний текст, відповідь відкритого типу, векторизація тексту, модель bag-of-words, TF, IDF, TF-IDF, TF-PWI, набір ознак для векторизації тексту.*

Kuzma K.T., Melnyk O.V. Research of methods for text vectorization in the tasks of validation the answers presented in natural language

Intellectualization of the process of processing natural language texts in the tasks of automated testing determines the relevance of the research. Since open-type answers in testing systems are natural language texts, the problem of their processing refers to the applied problem of word processing. All applied problems of word processing, the solution of which takes place

with the use of machine learning, neural networks, require vectorization – the conversion of text into digital values. The aim of the article is to research the models, methods of vectorization of texts in the problems of processing answers given in natural language.

At the first stage, the basic applied problems of word processing are investigated, as a result of which their classification is given. The assignment of the problem of checking natural language answers within the framework of this research to the problem of text classification and semantic analysis is substantiated.

In the second stage, the basic models of text representation in digital form are analyzed: bag-of-words and distributive semantics. The application of the bag-of-words model for the problem of processing open-ended answers is substantiated, as the vocabulary used to encode the collection of correct answers and the frequency of words with which they are used in the answers of "training" and "test" sets are enough to determine the answer class. It is noted that the vector of features in this problem is the frequency of tokens (symbolic or verbal uni-, bi-, n-grams) of the dictionary, formed by the training sample, in the answers of the "training" and "test" data sets.

In the third stage, the approaches of calculating the vector of characteristics are investigated: absolute frequency (TF), relative frequency (TF-IDF), compatible information (PWI), the advantages and disadvantages of each of them are determined.

At the last stage for vectorization of texts in problems of processing of the answers given in natural language, the following combinations of sets of signs are offered: model bag-of-words and TF; bag-of-words and TF-IDF model; verbal n-grams and TF-IDF; symbol n-grams and TF-IDF; model bag-of-words and TF-PWI.

The proposed sets of features and their combinations are a means of improving the machine learning model for the task of checking the answers given in natural language. Further research will be aimed at developing a model of machine learning of this problem and its experimental testing with the proposed sets of features in order to obtain an effective mathematical model.

Key words: *answer given in text form, natural language text, open-type answer, text vectorization, bag-of-words model, TF, IDF, TF-IDF, TF-PWI, set of features for text vectorization.*

Застосування технологій дистанційного навчання, цифровізація освітнього процесу зумовлюють актуальність дослідження процесів інтелектуалізації обробки природномовних текстів у задачах автоматизованої перевірки рівня засвоєння знань. Оскільки відповіді відкритого типу в системах тестування є природномовними текстами, то завдання їх обробки належить до прикладної задачі обробки текстової інформації.

Прикладними задачами (ті, які можна включити в основу розробки програмного продукту) обробки текстів є:

1) Класифікація за тематикою, тональністю [1–3]:
   а. довгих текстів;
   б. коротких текстів.
2) Пошук [4–5]:
   а. дублікатів документів;
   б. за запитом;
   в. тексту за зображенням або зображення за текстом;
   г. питально-відповідний пошук (QA).
3) Семантичний аналіз: STS (Semantic Textual Similarity) – семантична схожість текстів та виокремлення текстового підтексту Textual Entailment (TE) або Natural Language Inference (NLI) з метою визначення подібності та включення фрагментів тексту. STS класифікує за шкалою від 1 до 5 рівень семантичної еквівалентності між реченнями. RTE класифікує включення речень «так/ні» [6; 7].
4) Вилучення структурованої інформації (новини, медичні карти, електронні документи) [8].
5) Машинний переклад.
6) Діалогові системи (чат-боти).

Базовими підходами до вирішення зазначених задач є машинне навчання, нейромережі. Задача обробки відповідей, поданих природною мовою, може належати

як до категорії класифікації, семантичного аналізу, так і до задач питально-відповідного пошуку. В QA-системах запитання ставить користувач, а задачею системи є пошук релевантної відповіді з банку відповідей. Задачею ж валідації відповідей, поданих природною мовою, є перевірка наданої користувачем відповіді зі з'ясуванням відповідності заданому еталону. Тому задачу перевірки природномовних відповідей в межах цього дослідження віднесено до задачі класифікації та семантичного аналізу.

Всі прикладні задачі обробки природномовних текстів вимагають представлення тексту в цифровому вигляді – векторизації.

Мета статті – дослідити моделі, методи векторизації текстів у задачах обробки відповідей, поданих природною мовою.

Для вирішення задачі векторизації на першому етапі визначається модель представлення тексту. Вибір здійснюється між двома базовими моделями: модель bag-of-words і модель дистрибутивної семантики.

Bag-of-words – це модель, в якій текст (це може бути одне речення або весь текст) подається у вигляді безлічі слів. При цьому враховується кількість слів без урахування їхнього порядку та граматики. Таким чином, у моделі bag-of-words ознакою для навчання класифікатора є частота входження слова в тексті. У моделі дистрибутивної семантики враховується зв'язок слів між собою. Слова характеризуються своїм типовим контекстом (сусідніми словами). Ключовим об'єктом, який характеризує контекст, є матриця сумісного входження слів. Методи дистрибутивної семантики намагаються визначити смисл слів, аналізуючи розподіл ймовірностей входження слів у межах одного фрагмента тексту (або ймовірність зустріти одні слова в контексті інших). Модель дистрибутивної семантики є більш ресурсозалежною та вимагає більше часу на обробку даних. Оскільки для визначення класу відповіді досить складу словника, який застосовується для кодування колекції правильних відповідей, та частоти слів, з якою вони застосовуються у відповідях «навчального» і «тестового» наборів даних, то моделлю представлення тексту для задачі обробки відповідей відкритого типу обрано модель bag-of-words.

Вектор ознак у цій задачі – матриця, рядки якої відповідають відповідям, а стовпці – це токени (символьні або словесні уні-, бі-, n-грами), сформовані за навчальною вибіркою. На перетині рядка та стовпця вказується частота появи токена.

Найпростіший варіант – зважувати слова за кількістю їх вживань у відповіді. Ваги слів – це просто цілі числа. Недоліки такого підходу: вага слова залежить від довжини тексту. У довгих відповідях слова мають більшу вагу, нібито вони більш значущі, але це не так. До того ж самі частотні слова – це сполучники, прийменники, займенники. Вони зустрічаються всюди, але абсолютно неінформативні та не є корисними для будь-яких завдань класифікації.

Першим способом вирішення підвищення інформативності частот є нормування вектора відповіді на його довжину (або за евклідовою нормою) [9]. Якщо впорядкувати слова за спаданням частоти їх вживання, то вийде графік, який відповідає розподілу Ціпфа – розподілу ймовірностей, що описує відносини частоти події (вісь ординат) та кількості подій із такою частотою (вісь абсцис) [10]. Аналізуючи графік розподілу Ціпфа [10], можна зробити два практичних висновки: по-перше, частотних слів дуже мало. Вони не надто інформативні, адже зустрічаються практично в усіх відповідях. А ось рідкісних слів дуже багато – якщо ми якесь рідкісне слово зустрічаємо у відповіді, то з впевненістю можемо сказати, до якої тематики воно належить. Але проблема в тому, що такі слова дуже рідкісні й тому ненадійні як фактори під час прийняття рішень. Отже, потрібно

дотримуватися балансу частотності та інформативності. Основна ідея полягає в тому, що чим частіше слово зустрічається у відповіді, тим більше воно характерне для цієї відповіді, тим краще описує її тематику. З іншого боку, чим рідше це слово зустрічається в корпусі, у вибірці відповідей, тим більш воно специфічне й інформативне. За цей баланс відповідають дві величини: TF та IDF [11].

TF (term frequency) – це частота токена у відповіді:

$$TF = \frac{WordCount(w,d)}{length(d)},$$

де $WordCount(w,d)$ – кількість разів появи точена w у відповіді d; $length(d)$ – довжина відповіді d.

IDF – це обернена частота появи токена у відповіді. Розмір колекції ділиться на кількість відповідей, в яких вживається слово. Таким чином, найбільшу вагу матиме слово, що зустрічається лише в одній відповіді.

$$IDF(w,c) = \frac{size(c)}{DocCount(w,c)},$$

де $size(c)$ – розмір колекції відповідей; $DocCount(w,d)$ – кількість появи відповідей d, в яких зустрічається токен w, у колекції c.

Таким чином, кожна відповідь характеризується вектором відносних частот TF-IDF усіх токенів:

$$TF - IDF\ (w,d,c) = TF(w,d).$$

TF-IDF – це спосіб зважування та відбору категоріальних ознак у задачах машинного навчання [11]. Перевага TF-IDF полягає в тому, що цей метод можна використовувати, не маючи міток класу, тобто в завданнях «навчання без вчителя».

На практиці TF або IDF логарифмують з метою зменшення дисперсії відносних частот:

– логарифмування TF, коли тексти різної довжини:

$$\log(TF + 1).$$

– логарифмування IDF:

$$IDF(w,c) = \log\frac{size(c)}{DocCount(w,c)}.$$

Із урахуванням логарифмування IDF виходить, якщо токен часто з'являється в кожному тексті колекції, його IDF дорівнює 0; що рідше він зустрічається в корпусі, то більше IDF.

З метою попередження ділення на 0 виконують:

$$IDF(w,c) = \log\frac{size(c)}{DocCount(w,c)} + 1.$$

або «smooth»-логарифмування:

$$IDF(w,c) = \log\frac{size(c)+1}{DocCount(w,c)+1} + 1.$$

Останнім етапом підготовки текстових даних є стандартизація (масштабування) отриманих матриць, яка передбачає приведення значень отриманих частот до діапазону від 0 до 1. Існує кілька способів нормалізації даних для машинного навчання: мінімаксна нормалізація, центрування. Вибір залежить від конкретної задачі.

Є й інші способи вимірювання ознак за частотою – наприклад, взаємопов'язана інформація (Pointwise Mutual Information, PMI) [12–14]. Вона вимірюється між двома випадковими подіями або реалізацією двох випадкових величин. PMI

характеризує, наскільки сильно ми будемо очікувати першу подію, якщо перед цим спостерігаємо другу [14].

Для задачі класифікації відповідей:

$$pmi\,(l,w) = log\frac{p(l,w)}{p(l)\cdot p(w)} = log\frac{p(l\mid w)}{p(l)} = log\frac{p(w\mid l)}{p(w)},$$

де l – колекція відповідей, яка відповідає мітці класу L;

w – токен зі словника;

$p(w,l) = \frac{DocCount(w,l)}{Size(l)}$ – ймовірність зустріти токен w у відповіді класу L;

$p(w) = \frac{\sum_l DocCount(w,l)}{\sum_l Size\,(l)}$ – маргінальна ймовірність появи точена w;

$p\,(l) = \frac{Size\,(l)}{\sum_{,m} Size\,(m)}$ – маргінальна ймовірність зустріти відповідь класу L.

PMI (l,w) не залежить від однієї конкретної відповіді – лише від розподілів. Отже, PMI – це більш інформативний аналог IDF. Таким чином, можна об'єднати PMI із TF, отримавши ознаку TF-PMI.

Отже, обчислення частоти появи токена у відповідях, поданих природною мовою, можливе з використанням декількох підходів: абсолютної частоти (TF), відносної частоти (TF-IDF), сумісної інформації (PWI, TF-PWI).

Для векторизації текстів у задачах обробки відповідей, поданих природною мовою, пропонуються такі набори ознак:

1. Модель bag-of-words та TF.
2. Модель bag-of-words та TF-IDF.
3. Словесні n-грами та TF-IDF (оскільки тексти відповідей є короткими, то розглядаються 1,2-грами та їх поєднання).
4. Символьні n-грами (n: 1…6) та TF-IDF.
5. Модель bag-of-words та TF-PWI.

Запропоновані набори ознак та їх комбінації є засобами покращення моделі машинного навчання для задачі перевірки відповідей, поданих природною мовою. Подальші дослідження будуть спрямовані на експериментальне тестування запропонованих наборів характеристик під час обробки відповідей, поданих природною мовою, визначення того набору або комбінації, що забезпечить найкращі показники валідації відповідей із використанням методів машинного навчання.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ:

1. Zhang L.J. et al. Deep learning for sentiment analysis: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2018. DOI: 10.1002/widm.1253.

2. Лесько О.М., Рогушина Ю.В. Использование онтологий для анализа семантики естественно-языковых текстов. *Проблеми програмування.* 2009. № 3. С. 59−65.

3. Ваколюк Т.В., Комарницька О.І. Алгоритм нечіткого семантичного порівняння текстової інформації. *Збірник наукових праць Військового інституту Київського національного університету ім. Т. Шевченка*. 2013. № 39. С. 163−168.

4. Цыганов Н.Л., Циканин М.А. Исследование методов поиска дубликатов веб-документов с учетом запроса пользователя. *Интернет-математика-2007* : сборник работ участников конкурса. 2007. Екатеринбург : Издательство Уральского университета. С. 211–222.

5. Mutabazi E. et al. A Review on Medical Textual Question Answering Systems Based on Deep Learning Approaches. Applied Sciences. 2021. No. 11 (12). DOI: 10.3390/app11125456.

6. Rocktäschel T. et al. Reasoning about entailment with neural attention. *arXiv preprint arXiv*.1509.06664. 2015.

7. Ampomah I.K., Park S.B., Lee S.J. A Sentence-to-Sentence Relation Network for Recognizing Textual Entailment. *World Academy of Science, Engineering and Technology International Journal of Computer and Information Engineering*. 2016. Nov. 1; 10 (12): 1955-8.

8. Годич О.В., Наконечний Ю.С., Щербина Ю.М. Категоризація електронних документів. *Вісник Національного університету «Львівська політехніка» «Інформаційні системи та мережі».* 2010. № 673. С. 233–248.

9. Euclidean norm. *Wikipedia: the free encyclopedia*. URL: https://en.wikipedia.org/wiki/Norm_(mathematics) (дата звернення: 15.09.2021 р.).

10. Zipf's law. *Wikipedia: the free encyclopedia*. URL: https://en.wikipedia.org/wiki/Zipf%27s_law (дата звернення: 15.09.2021 р.).

11. TF-IDF. *Wikipedia: the free encyclopedia*. URL: https://en.wikipedia.org/wiki/Tf-idf (дата звернення: 20.11.2021 р.).

12. Pointwise mutual information. *Wikipedia: the free encyclopedia*. URL: https://en.wikipedia.org/wiki/Pointwise_mutual_information (дата звернення: 20.11.2021 р.).

13. Mutual information. *Wikipedia: the free encyclopedia*. URL: https://en.wikipedia.org/wiki/Mutual_information (дата звернення: 20.11.2021 р.).

14. Levy O., Goldberg Yoav. Neural Word Embedding as Implicit Matrix Factorization. *Advances in neural information processing systems*. 2014. № 27, pp. 2177–2185.

REFERENCES:

1. Zhang, L., Wang, S., Liu, B. (2018) Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery.* № 8 (4), e1253.

2. Les'ko, O.M., Rogushina, Y.U. (2009) Ispol'zovanie ontologij dlya analiza semantiki estestvenno-yazykovyh tekstov [The usage of ontologies for semantics analysis of texts on natural language]. *Problemi programuvannya*, (3), pp. 59−65. [in Russian]

3. Vakoliuk, T.V., Komarnytska, O.I. (2013) Alhorytm nechitkoho semantychnoho porivniannia tekstovoi informatsii. *Zbirnyk naukovykh prats Viiskovoho instytutu Kyivskoho natsionalnoho universytetu im. T. Shevchenka*, (39), pp. 163−168. [in Ukrainian]

4. Cyganov, N.L., Cikanin, M.A. (2007) Issledovanie metodov poiska dublikatov veb-dokumentov s uchetom zaprosa pol'zovatelya [Investigating techniques of fuzzy duplicate web-documents detection based on a user's request]. *Internet-matematika-2007.* Ekaterinburg. [in Russian]

5. Mutabazi, E., Ni, J., Tang, G., Cao, W. (2021) A Review on Medical Textual Question Answering Systems Based on Deep Learning Approaches. Applied Sciences. No. 11, p. 5456.

6. Rocktäschel, T., Grefenstette, E., Hermann, K.M., Kočiský, T., Blunsom, P. (2015) Reasoning about entailment with neural attention. *arXiv preprint arXiv.* 1509.06664.

7. Ampomah, I.K., Park, S., Lee, S. (2016) A Sentence-to-Sentence Relation Network for Recognizing Textual Entailment. World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering. No. 10*,* pp. 2060–2063.

8. Hodych, O.V., Nakonechnyi, Yu.S., Shcherbyna, Yu.M. (2010) Katehoryzatsiia elektronnykh dokumentiv. *Visnyk Natsionalnoho universytetu "Lvivska politekhnika" "Informatsiini systemy ta merezhi"*. No. 673, pp. 233–248. [in Ukrainian]

9. Euclidean norm. *Wikipedia.org.* URL: https://en.wikipedia.org/wiki/Norm_(mathematics).
10. Zipf's law. *Wikipedia.org.* URL: https://en.wikipedia.org/wiki/Zipf%27s_law.
11. TF-IDF. *Wikipedia.org.* URL: https://en.wikipedia.org/wiki/Tf-idf.
12. Pointwise mutual information. *Wikipedia.org.* URL: https://en.wikipedia.org/wiki/Pointwise_mutual_information.
13. Mutual information. *Wikipedia.org.* URL: https://en.wikipedia.org/wiki/Mutual_information.
14. Levy, O., Goldberg, Y. (2014) Neural Word Embedding as Implicit Matrix Factorization. *NIPS.*