УДК 004.512.4

О. В. БІСІКАЛО, С. М. ДОВГАЛЕЦЬ, А. І. ЛІСОВЕНКО

СТВОРЕННЯ ДІАЛОГОВОЇ СИСТЕМИ ДЛЯ ТЕКСТОВОГО НАВЧАЛЬНОГО КОНТЕНТУ

Вінницький національний технічний університет,
21021, Хмельницьке шосе 95, м. Вінниця, Україна
тел.+38(093)18-68-399, E-mail: sweettee@ukr.net

Анотація. Статтю присвячено актуальній тематиці створення діалогових систем, які здатні давати відповіді користувачам в залежності від вміщеного до них текстового навчального контенту. Для підтримки діалогу запропоновано модель бази знань на основі одиниці та нечіткого відношення сенсу. Лексичні знання вилучаються з тексту шляхом оброблення синтаксичних зв'язків між словоформами всіх його речень, а відповідь генерується як композиція нечітких відношень за формальним критерієм. Інформаційну технологію підтримки діалогу апробовано на тестовому прикладі, який демонструє у відповідях поєднання інформації з кількох речень.

Ключові слова: інформаційна технологія, діалогова система, текстовий навчальний контент, сила зв'язків, словоформи, нечітке відношення, композиція.

Аннотация. Статья посвящена актуальной тематике создания диалоговых систем, которые способны давать ответы пользователям в зависимости от помещенного в них текстового учебного контента. Для поддержки диалога предложена модель базы знаний на основе единицы и нечеткого отношения смысла. Лексические знания извлекаются из текста путем обработки синтаксических связей между словоформами всех его предложений, а ответ генерируется как композиция нечетких отношений по формальным критериям. Информационная технология поддержки диалога апробирована на тестовом примере, который демонстрирует в ответах сочетание информации из нескольких предложений.

Ключевые слова: информационная технология, диалоговая система, текстовый учебный контент, сила связи, словоформы, нечеткое отношение, композиция.

Abstract. Article is devoted to actual scope of creation of dialogue systems which are capable to give answers to users depending on the text educational content placed in them. For maintenance of dialogue the knowledge base model on the basis of unit and the indistinct relation of sense is offered. Lexical knowledges is extracted from the text by processing of syntactic links between word forms of all its sentences, and the answer is generated as composition of the fuzzy relations by formal criteria. Information technology of maintenance of dialogue are approved on a test example which shows in answers the combination of information from several sentences.

Keywords: information technology, dialogue system, text educational content, relations force, word forms, fuzzy relation, composition.

ВСТУП

В умовах світу, що розвивається, значне місце посідає пошук інформації та надання відповіді на конкретні питання користувача у певній предметній області, що його цікавить. Як наслідок, все більшої актуальності набувають системи, що мають здатність до самонавчання — зокрема інтелектуальні діалогові системи, які генерують таку відповідь на запитання користувача у певній області, що максимально наближена до відповіді людини-експерта. Потребують розроблення універсальні системи навчального спрямування, які, маючи на вході певний навчальний текст, перетворюють його на таку базу

знань відповідної предметної області, що забезпечує відповідь на питання не тільки одним реченням (як пошукові системи), але й комбінованою інформацією з кількох речень (за прикладом людини).

Метою дослідження є розроблення математичного забезпечення діалогової системи та апробація відповідної інформаційної технології, що має здатність:

— перетворювати навчальний контент у вигляді вхідної текстової інформації на формальну базу знань;

— представляти питання користувача системи (учня) у форматі, сумісному з базою знань;

— генерувати відповідь на питання як згортку елементів знань за формальним критерієм;

— інтерпретувати отриману відповідь у формі природно-мовної конструкції, що нагадує речення.

МОДЕЛЬ БАЗИ ЗНАНЬ НА ОСНОВІ ОДИНИЦІ СЕНСУ

Певним обмеженням на першому етапі розробки системи вважатимемо генерацію відповідей лише у межах речень одного навчального тексту. Для реалізації поставленої мети пропонується використовувати нечіткі відношення.

Вхідний текст представимо у вигляді матриці сили зв’язків між словоформами у тексті. Такий результат є наслідком синтаксичного розбору кожного речення у вхідному тексті. Пропонується накопичувати такі зв'язки для всіх речень тексту у вигляді матриці Q, в якій рядки (головне слово пари) та стовпці (підлегле слово) відповідають тим словоформам, що повторюються в різних реченнях. Особливість підходу полягає у врахуванні сили зв’язків лише між значимими словоформами. Отримати такого роду відформатовану матрицю можна вручну (для невеликих текстів), що забезпечує абсолютну точність синтаксичного аналізу, а також за допомогою програмних лінгвістичних технологій, наприклад, DKPro [1] — у цьому випадку потрібно розраховувати на визначення зв’язків з похибкою до 10 %.

Припустимо, що навчальний вхідний текст складається з L речень, тоді для його відображення у нечітке відношення побудуємо функцію належності для створеної нечіткої множини значимих словоформ. З цією метою застосуємо функцію належності на зразок запропонованої у [2] чисельної міри сенсу 1 *Сав*. Зокрема значення функції належності будемо вважати одиницею сенсу розміром

$$\mu_Q(< i_l, i_j >) = 1,$$

де значення елемента *(l,j)* матриці Q залежить від статистики появи зв’язку для кортежу $< i_l, i_j >$ за час спостереження (аналізу) L вхідних речень.

В загальному вигляді функція належності нечіткого відношення сенсу для пар зі словоформ задається як

$$\mu_Q(< i_l, i_j >) = f(k_{lj}, t_L), \tag{1}$$

де k_{lj} — кількість зафіксованих зв’язків між l-им та j-им елементами створеної матриці Q в момент часу t_L.

Для визначення рівня імовірнісного прогнозування нормуємо функцію належності у проміжку [0, 1]. Для цього розрахуємо статистичну оцінку λ (математичне сподівання): якщо $k_\Sigma = \sum_{i=1}^{n} \sum_{j=1}^{n} k_{lj}$ та m — кількість ненульових елементів матриці Q, то $\lambda = k_\Sigma / m$ – в цьому випадку застосовуємо відому сигмоїдальну функцію, а саме

$$\mu_Q(< i_l, i_j >) = f(k_{lj}, \lambda) = \frac{1}{1 + e^{-k_{lj}+\lambda}}. \tag{2}$$

де k_{lj} — всі ненульові елементи матриці.

Наслідком нормування є поява характеристичної властивості функції належності, отриманої внаслідок запропонованого підходу [2] — середнє значення нечіткого відношення сенсу пари словоформ навчального тексту дорівнює 0,5. Формально

$$\overline{\mu_Q} = \frac{1}{m} \sum_{j=1}^{m} \mu_{Qj} = 0{,}5.$$

Аналогічним чином побудуємо матрицю запитання користувача R. Для цього будемо використовувати як шаблон матрицю, що аналогічна матриці вхідного тексту Q — того ж розміру, але із зануленими комірками. За допомогою синтаксичного розбору питального речення вносимо зв’язки між словоформами у комірки з відповідними словоформами існуючого шаблону матриці R. Отже, ми маємо матрицю тексту Q та матрицю запитання R. Важливою умовою для подальшої обробки цих матриць є те,

що обидві вони повинні бути квадратними та одного розміру, але ця умова вводить додаткове обмеження моделі — питання можна задавати тільки з тих слів, що вже є наявними у навчальному тексті.

Наступним кроком закладемо визначення матриці *P*, яка надасть відповідь по вхідному тексту на запитання користувача. Для цього нам необхідно визначити функції належності матриці *Q* та функцію належності матриці *R* згідно рівняння (2). Наприклад, для певного текстового прикладу, матриці *Q* та *R* мають розмірність 164×164. Припустимо, що математичні сподівання матриць дорівнюють: $\lambda_Q = 1{,}116$ та $\lambda_R = 1$. Тоді, за формулою (2) отримаємо значення належностей μ_Q та μ_R для кожної пари $< i_l, i_j >$.

Для отримання відповіді застосуємо формули композиції нечітких відношень «MAX-MIN»

$$\mu_P\left(< i_l, i_j >\right) = \max_{i_k \in I}\{\min\{\mu_Q\left(< i_l, i_k >\right), \mu_R\left(< i_k, i_j >\right)\}\}. \quad (3)$$

Застосувавши дану формулу (3), ми отримаємо нову матрицю *P*, яка буде містити різні числові значення з $\mu_{\mathbf{P}}$ в тих комірках, які найімовірніше будуть відповіддю на запитання користувача.

Для порівняння та достовірності результатів використаємо ще один різновид композиції нечітких відношень «MIN-MAX» [3]

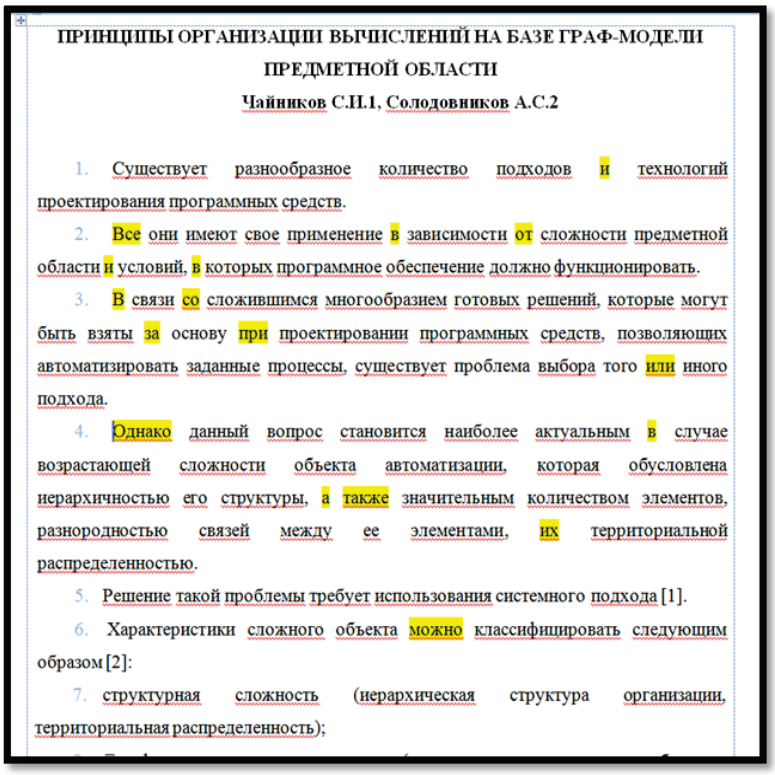
$$\mu_P\left(< i_l, i_j >\right) = \min_{i_k \in I}\left\{\max\left\{\mu_Q\left(< i_l, i_k >\right), \mu_R\left(< i_k, i_j >\right)\right\}\right\}. \quad (4)$$

За результатами роботи двох типів композиції нечітких відношень будемо очікувати, що система надасть такі відповіді, які мають сенс та задовольнять запитанню користувача.

Для визначення відповіді пропонується відсортувати за формальним критерієм список пар «головне та залежне слово», наприклад за силою зв'язків між відповідними словоформами. Тоді набір пар, які мають найбільше чисельне значення і перевищують певну задану межу сформує відповідь.

АПРОБАЦІЯ ДІАЛОГОВОЇ СИСТЕМИ НА ТЕСТОВОМУ ПРИКЛАДІ

Розглянемо приклад роботи моделі та відповідної інформаційної технології підтримки діалогу з навчальним контентом. Як тестовий приклад було довільно обрано російськомовний текст [4], з якого обрано для апробації діалогової системи перших 29 речень. На першому етапі в даному тексті кожне речення пронумеровано, з нього вибрано лише значимі частини мови, тому незначимі (не враховані) позначено кольором. На рисунку 1 наведено фрагмент синтаксичного розбору обраного тексту.



ПРИНЦИПЫ ОРГАНИЗАЦИИ ВЫЧИСЛЕНИЙ НА БАЗЕ ГРАФ-МОДЕЛИ ПРЕДМЕТНОЙ ОБЛАСТИ

Чайников С.П.1, Солодовников А.С.2

1. Существует разнообразное количество подходов и технологий проектирования программных средств.
2. Все они имеют свое применение в зависимости от сложности предметной области и условий, в которых программное обеспечение должно функционировать.
3. В связи со сложившимся многообразием готовых решений, которые могут быть взяты за основу при проектировании программных средств, позволяющих автоматизировать заданные процессы, существует проблема выбора того или иного подхода.
4. Однако данный вопрос становится наиболее актуальным в случае возрастающей сложности объекта автоматизации, которая обусловлена иерархичностью его структуры, а также значительным количеством элементов, разнородностью связей между ее элементами, их территориальной распределенностью.
5. Решение такой проблемы требует использования системного подхода [1].
6. Характеристики сложного объекта можно классифицировать следующим образом [2]:
7. структурная сложность (иерархическая структура организации, территориальная распределенность);

Рис. 1. Фрагмент тексту, що обрано для тестового прикладу

Всі обчислення для інформаційної технології було здійснено у математичному пакеті MatLab.

За результатами проведеного синтаксичного аналізу тексту було побудовано матрицю розмірністю 164×164 з визначеною силою зв'язків між словоформам, що входять до даного тексту. Також зазначимо, що сума всіх не порожніх елементів матриці Q дорівнює 319; кількість елементів матриці Q із силою зв'язку = 1 — 295; сума всіх зв'язків між словоформами у матриці Q — 356.

На рисунку 2 зображено типову діаграму залежності сили зв'язків словоформ матриці тексту Q, в якій вісь абсцис відповідає головній словоформі, а вісь ординат позначає силу її зв'язку k_{1j} із залежною словоформою. Більшість зв'язків — одиничні, а найвагоміші для тестового прикладу не переважають 6.

Для створення питального речення від користувача було обрано приклад: «*От чего зависит диалоговый режим работы системы?*». Для отримання відповіді системою проведемо синтаксичний розбір даного питального речення та внесемо зв'язки між словоформами у пусту шаблонну матрицю R. Дана матриця також матиме розмір 164×164 та утворюється шляхом занулення усіх комірок матриці Q. На рисунку 3 зображено синтаксичний розбір питального речення, в якому враховуються лише значимі частини мови.

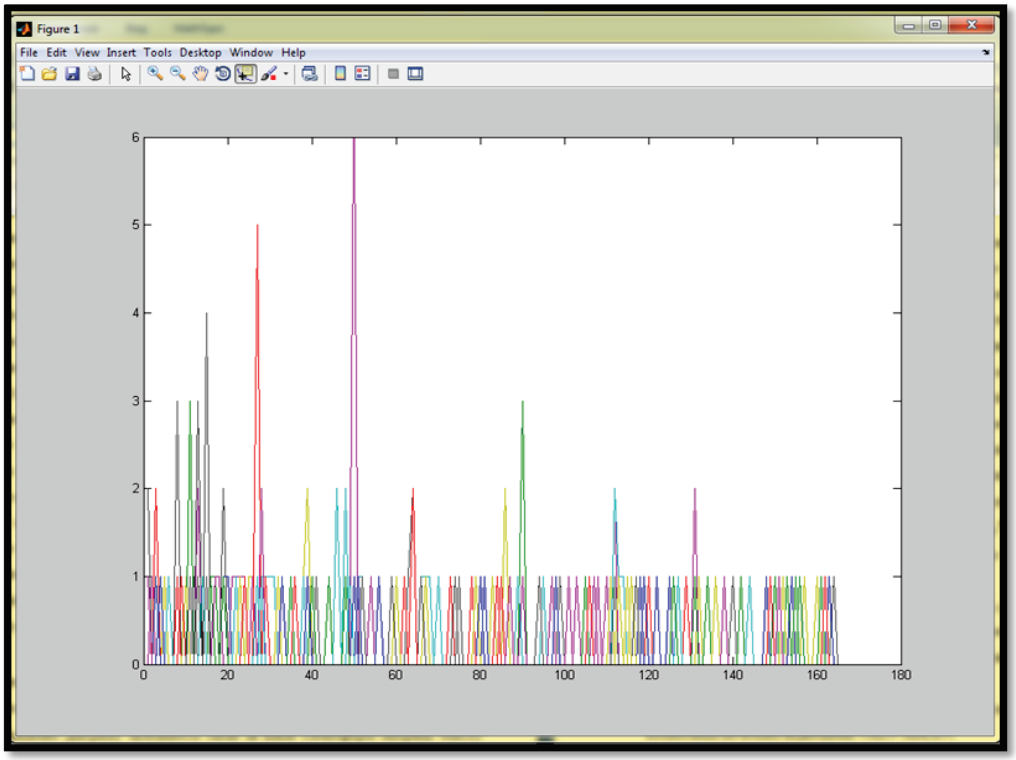


Рис. .2. Діаграма залежності сили зв'язків матриці Q між словоформами

Рис. 3. Синтаксичний розбір питального речення користувача

На рисунку 4 представлено діаграму сили зв'язку матриці R між словоформами у питальному реченні, в якій вісі позначено як — X (головна словоформа) та Y (сила її зв'язку згідно (2) із залежною).

Скориставшись рівняннями (1) та (2), та перетворивши матриці *Q* та *R* у набори нечітких відношень, використаємо рівняння (3). Це рівняння є максимінною композицією нечітких відношень [3]. В результаті застосування даної формули отримаємо нову матрицю *P*, яка в порожню шаблонну матрицю *Q* внесе нові відомості для подальшого визначення відповіді.

На рисунку 5 представлено діаграму залежності сили зв'язків між парами словоформ при використанні «MAX-MIN» композиції відношень (позначення вісей аналогічні рисунку 4), а у таблиці 1 ці результати проінтерпретовано у вигляді словосполучень, що відповідають обмеженням за силою.

Отже, бачимо, що результати композиції із більшою межею сили зв'язку дають більш лаконічну та точну відповідь, що інтерпретується: *«От управления системных зависимостей и управления диалоговой системы»*.

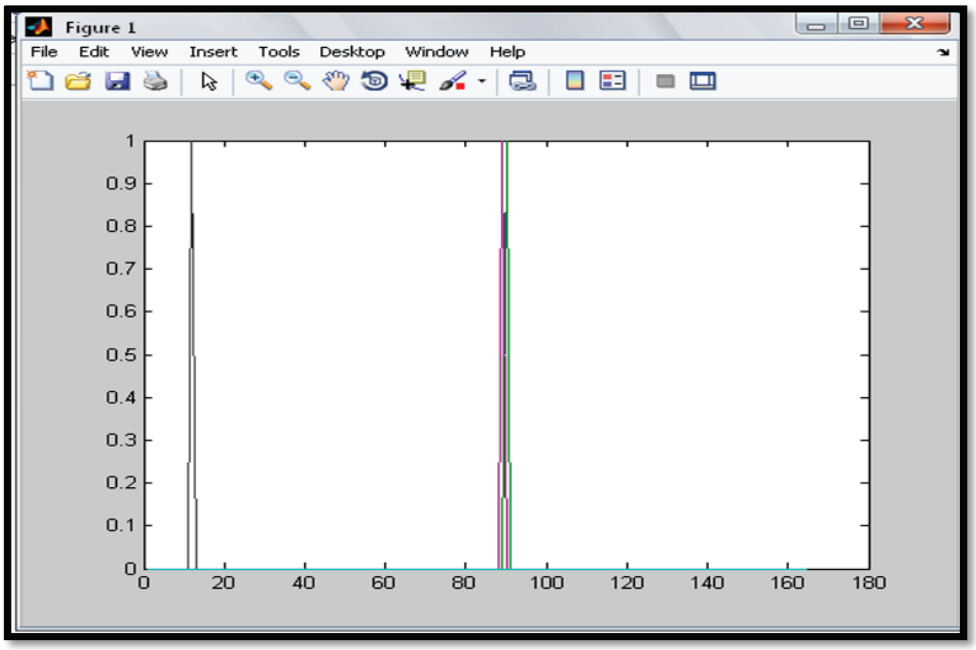


Рис. 4. Діаграма залежності сили зв'язків матриці *R* між словоформами

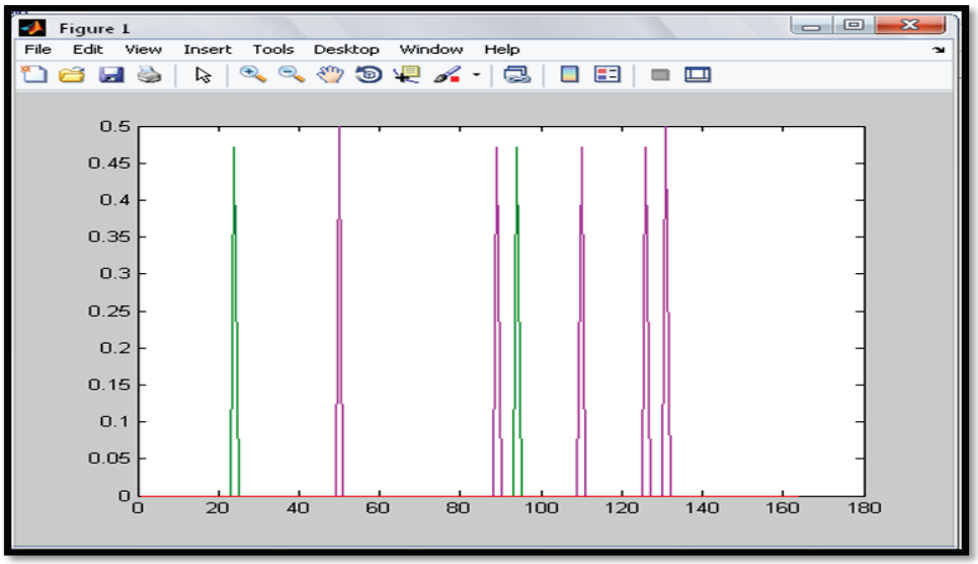


Рис. 5. Діаграма залежності сили зв'язків матриці *P* між словоформами («MAX-MIN»)

Таблиця 1.

Результати виконання «MAX-MIN» композиції відношень

№п/п	0,5	0,4709
1	системной зависимости	общение зависимости
2	управляет зависимостью	позволяющих систему
3	диалоговая система	общение диалога
4	управляет диалогом	отношение диалога
5		режим зависимости
6		режим диалога
7		оптимизировать систему
8		отношение зависимости

З метою порівняння застосуємо ще один вид композиції нечітких відношень, а саме, мінімаксну. Для цього використаємо рівняння (4) та проведемо аналогічні розрахунки з нечіткими відношеннями матриць *Q* та *R*. На рисунку 6 представлено діаграму залежності сили зв'язків між словоформами при використанні «MIN-MAX» композиції, позначення вісей аналогічні рисункам 4 та 5.

У таблиці 2 наведено та відповідно проінтерпретовано результати роботи системи за «MIN-MAX» композицією.

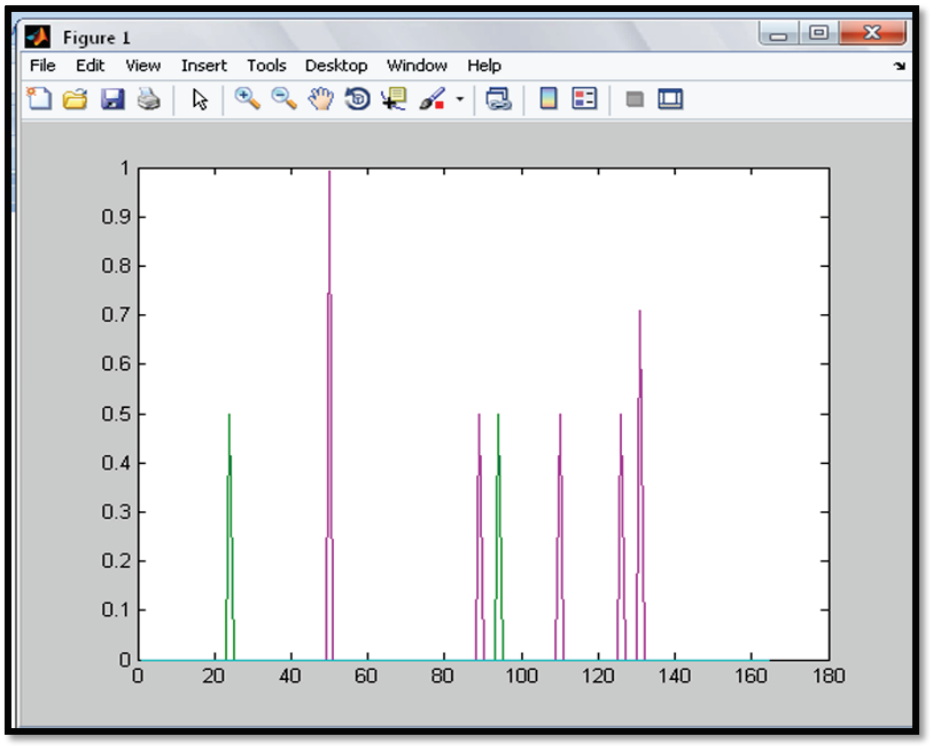


Рис. 6. Діаграма залежності сили зв'язків матриці *P* між словоформами («MIN-MAX»)

Таблиця 2.

Результати виконання «MIN-MAX» композиції відношень

№п/п	0,9925	0,7077	0,5
1	системной зависимости	управляет зависимостью	позволяет систему
2	системного диалога	управляет диалогом	режим зависимости
3			режим диалога
4			оптимизировать систему
5			отношению зависимости
6			отношению диалога
7			общение зависимого
8			общение диалога

За найбільшою силою зв'язків між словоформами знову отримано як найлаконічніші перші два варіанти. Але тепер інтерпретацією можливих відповідей є *«От системной зависимости и системного диалога»* або *«От управления зависимостью и диалогом»*. Помітно, що дані відповіді «MIN-MAX» менш якісні за змістом, аніж у першому варіанті «MAX-MIN», проте і ці відповіді не позбавлені певного сенсу.

ВИСНОВКИ

Запропонований в роботі підхід до створення діалогових систем навчального спрямування побудовано за допомогою вилучення лексичних знань з речень тексту та окремих природно-мовних конструкцій питального типу на основі одиниці сенсу. Внаслідок цього забезпечено можливість отримати відповідь на питання не тільки одним реченням навчального тексту, але й комбінованою інформацією з кількох речень. Обмеження моделі на даному етапі дослідження полягає у необхідності задавати питання тільки з тих слів, що вже є наявними у навчальному тексті

Отримано та апробовано на тестовому прикладі інформаційну технологію підтримки діалогової системи, що генерує відповідь на питання як результат композиції нечітких відношень сенсу. Виявлено, що найбільш лаконічну та точну відповідь забезпечують множини словосполучень, які відповідають парам словоформ з найвищими межами ваги. При цьому найкращі відповіді за «MAX-MIN» композицією переважають найкращі за «MIN-MAX» композицією за якістю змісту.

Перспективним напрямом подальших досліджень необхідно вважати проведення масштабних експериментів з великими текстами навчального спрямування та різними категоріями питань, ознаками яких є питальні займенники.

СПИСОК ЛІТЕРАТУРИ

1. Natural Language Processing: Integration of Automatic and Manual Analysis [Electronic resource]. — Technischen Universität Darmstadt, 2014. — Available at: \www/URL: http://tuprints.ulb.tu-darmstadt.de/4151/1/rec-thesis-final.pdf. — 21.04.2015.
2. Бісікало О. В. Формальні методи образного аналізу та синтезу природно-мовних конструкцій : монографія [Текст] / О. В. Бісікало // — Вінниця : ВНТУ, 2013. — 316 с. — ISBN 978-966-641-528-1.
3. Штовба С. Д. Введение в теорию нечетких множеств и нечеткую логику / С. Д. Штовба // [Електронный ресурс]. — Доступ: \www/URL: http://matlab.exponenta.ru/fuzzylogic/book1/1_5.php.
4. Чайников С. И. принципы организации вычислений на базе граф-модели предметной области / С. И Чайников, А. С. Солодовников // [Електронный ресурс]. — Доступ: \www/URL:http://repo.knmu.edu.ua/bitstream/123456789/3099/1/%D1%81%D1%821%D0%B1%D0%B8%D0%BE%D0%BD%D0%B8%D0%BA%D0%B0%D0%B8%D0%BD%D1%82.pdf

SPISOK LITERATURI

1. Natural Language Processing: Integration of Automatic and Manual Analysis [Electronic resource]. — Technischen Universität Darmstadt, 2014. — Available at: \www/URL: http://tuprints.ulb.tu-darmstadt.de/4151/1/rec-thesis-final.pdf. — 21.04.2015.
2. Bisikalo O. V. Formalni metodi obraznogo analizu ta sintezu prirodno-movnih konstruktsiy : monografiya [Tekst] / O. V. Bisikalo // — Vinnitsya : VNTU, 2013. — 316 s. — ISBN 978-966-641-528-1.
3. Shtovba S. D. Vvedenie v teoriyu nechetkih mnozhestv i nechetkuyu logiku / S. D. Shtovba // [Elektronnyiy resurs]. — Dostup: \www/URL: http://matlab.exponenta.ru/fuzzylogic/book1/1_5.php.
4. Chaynikov S. I. printsipyi organizatsii vyichisleniy na baze graf-modeli predmetnoy oblasti / S. I Chaynikov, A. S. Solodovnikov // [Elektronnyiy resurs]. — Dostup: \www/URL: http://repo.knmu.edu.ua/bitstream/123456789/3099/1/%D1%81%D1%821%D0%B1%D0%B8%D0%BE%D0%BD%D0%B8%D0%BA%D0%B0%D0%B8%D0%BD%D1%82.pdf

Надійшла до редакції 23.11.2015 р.

БІСІКАЛО ОЛЕГ ВОЛОДИМИРОВИЧ — д.т.н., професор кафедри АІВТ, декан ФКСА, ВНТУ, м.Вінниця, Україна.

ДОВГАЛЕЦЬ СЕРГІЙ МИХАЙЛОВИЧ — к.т.н., доцент кафедри АІВТ, ВНТУ, м.Вінниця, Україна.

ЛІСОВЕНКО АННА ІГОРІВНА — аспірант кафедри АІВТ, ВНТУ, м.Вінниця, Україна.