Міністерство освіти і науки України
Київський національний університет імені Тараса Шевченка

ВОЗНЮК Тарас Григорович

УДК 681.3.062

ЗАСТОСУВАННЯ СЕМАНТИКО-СИНТАКСИЧНОЇ ТЕНЗОРНОЇ МОДЕЛІ ПРИРОДНОЇ МОВИ ДЛЯ АНАЛІЗУ КОРЕФЕРЕНТНИХ ЗВ'ЯЗКІВ У ТЕКСТАХ

01.05.01 – теоретичні основи інформатики та кібернетики

Автореферат

дисертації на здобуття наукового ступеня
кандидата фізико-математичних наук

Київ – 2016

Дисертацією є рукопис.

Робота виконана на кафедрі математичної інформатики факультету кібернетики Київського національного університету імені Тараса Шевченка Міністерства освіти і науки України.

Науковий керівник: доктор фізико-математичних наук, доцент
Марченко Олександр Олександрович,
Київський національний університет імені Тараса Шевченка, доцент кафедри математичної інформатики

Офіційні опоненти: доктор фізико-математичних наук, професор
Скобелєв Володимир Геннадійович,
Інститут кібернетики імені В.М. Глушкова НАН України, м. Київ
провідний науковий співробітник

кандидат фізико-математичних наук, доцент
Глибовець Андрій Миколайович,
Національний університет «Києво-Могилянська Академія», м. Київ
доцент кафедри мережних технологій

Захист відбудеться 24 березня 2016 р. о 14 годині на засіданні спеціалізованої вченої ради Д 26.001.09 Київського національного університету імені Тараса Шевченка за адресою: 03680, Київ, проспект Академіка Глушкова, 4д, факультет кібернетики, ауд. 40.

З дисертацією можна ознайомитись у Науковій бібліотеці ім. М.Максимовича Київського національного університету імені Тараса Шевченка за адресою: 01601 МСП, Київ, вул. Володимирська, 58.

Автореферат розісланий “_____” лютого 2016 р.

Вчений секретар
спеціалізованої вченої ради В.П. Шевченко

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність теми дослідження. В наш час бурхливого розширення сфер застосування інформаційних технологій задачі обробки текстів природною мовою набувають великого значення в науці, економіці та інших сферах життя суспільства. Автоматичний переклад тексту допомагає людям з різних країн розуміти один одного без докладання значних зусиль та витрат часу. Компанії-гіганти, такі як Google та Facebook застосовують методи комп'ютерної обробки текстів для покращення точності цільової реклами, аналізуючи переписку та пошукові запити користувачів. Фірми, що займаються соціологічними дослідженнями, мають змогу оцінювати ставлення людей до певних осіб чи подій на основі автоматичного аналізу публікацій у пресі, залишених користувачами коментарів, повідомлень на форумах. І це далеко не повний перелік практичних задач, для розв'язання яких використовуються методи та алгоритми комп'ютерної лінгвістики.

Для того, щоб автоматичний переклад текстів одразу видавав результат, який не потребує кропіткої додаткової вичитки, необхідно, щоб комп'ютер був здатний розуміти текст на рівні людини. Для оцінки складності задач штучного інтелекту аналогічно до класу NP-повних задач, введено клас AI-повних (**A**rtificial **I**ntelligence) задач, повне вирішення яких вимагає побудови штучного інтелекту, близького до рівня інтелектуальних здібностей дорослої людини.

Задача повного розуміння текстів є вкрай складною, тому її розбивають на підзадачі – розуміння значень окремих слів, розуміння певних типів зв'язків між словами та їх залежностей, розуміння речень природною мовою в контексті всього текста в цілому чи в контексті знань людини про конкретну предметну область. При автоматичному аналізі природномовних текстів, необхідно різні слова ототожнювати з однією сутністю. Ця підзадача важлива для більш глибокого розуміння тексту. Зв'язок що виникає між такими словами називається кореферентним. Він виникає між словами та словосполученнями, що посилаються на один і той самий об'єкт позамовної дійсності. Так як іменник "дівчина" та займенник "вона" в загальному випадку не обов'язково вказують на одну і ту саму людину, наявність такого зв'язку можна визначити лише з контексту в якому слова були вжиті. Вирішенню задачі кореферентного аналізу були присвячені роботи Шолома Лапіна, Герберта Ліса, Руслана Міткова, Хійан Лі, Анатолія Анісімова, Олександра Марченка та багатьох інших вчених.

Для аналізу текстів довільної тематики традиційні підходи зі складанням словників та правил їх обробки вимагають багато кропіткої праці. Саме тому актуальною є розробка систем, що здатні до самонавчання із автоматизованим виділенням даних про навколишній світ з нерозмічених природномовних текстів. Для вирішення даної задачі можуть бути використані такі потужні моделі як керуючі простори синтаксичних структур речень природною мовою,

реалізовані за допомогою методів невід'ємної факторизації лінгвістичних тензорів з елементами машинного навчання. Це в свою чергу вимагає розвитку зазначених моделей семантико-синтаксичних структур природної мови та алгоритмів їх обробки. Важливість розв'язання задачі автоматизації побудови кореферентних зв'язків при обробці природномовних текстів і визначає актуальність дисертаційної роботи.

Зв'язок роботи з науковими програмами, планами, темами. Дисертаційна робота є складовою частиною наукових робіт, які ведуться на кафедрі математичної інформатики факультету кібернетики Київського національного університету імені Тараса Шевченка при виконанні фундаментальної теми "Створення теоретичних основ методів та програмних засобів інтелектуалізації інформаційно–комунікаційних та трансформерних технологій" (державний номер реєстрації – 0111U005416, 2011–2015 рр.)

Мета і задачі дисертаційного дослідження. Метою дисертаційної роботи є побудова математичних моделей представлення семантико-синтаксичних структур текстів та розробка алгоритму кореферентного аналізу на основі створених моделей.

З огляду на мету в роботі ставляться такі задачі:

1. Розробити нові методи оцінки наявності кореферентного зв'язку між парою сутностей за допомогою методів машинного навчання на розширеному семантико-синтаксичними ознаками просторі.
2. Побудувати та дослідити алгоритми оцінки синтаксичного та семантичного паралелізму на основі тензорної моделі.
3. Розробити та математично обґрунтувати алгоритм побудови багатовимірного тензору опису структур природної мови та архітектуру системи, що здатна обробляти великі текстові корпуси.
4. Розробити алгоритм побудови керуючих просторів синтаксичних структур для збагачення тензорної моделі мови та довести його коректність та обчислити складність в термінах швидкодії та пам'яті.
5. Провести експерименти з оцінки точності роботи алгоритму знаходження кореферентних зв'язків та порівняти його результати з іншими алгоритмами.

Об'єкт дослідження – моделі семантико-синтаксичних структур природномовних текстів та методи аналізу текстів на їх основі.

Предмет дослідження – алгоритми кореферентного аналізу в контексті семантико-синтаксичних тензорних моделей.

Методи дослідження. Дослідження базуються на методах та алгоритмах теорії графів, теорії синтаксичного аналізу, тензорного числення, штучного інтелекту, машинного навчання, методики побудови комп'ютерно-лінгвістичних систем.

Наукова новизна одержаних результатів. У дисертаційній роботі розроблено та математично обґрунтовано алгоритми для вирішення задачі ідентифікації та аналізу кореферентних зв'язків у природномовних текстах і отримано такі нові наукові результати:

1. Розроблено нові методи оцінки наявності кореферентного зв'язку між парою слів за допомогою машинного навчання.
2. В методі опорних векторів удосконалено алгоритм навчання для класифікації кореферентних сутностей. Це дало змогу одержати більш точні результати класифікації для типової задачі знаходження кореферентностей, коли кількість негативних прикладів на декілька порядків перевищує кількість позитивних прикладів.
3. Для підвищення точності роботи класифікатора було розроблено розширений простір ознак із додаванням семантико-синтаксичних ознак.
4. Для обчислення нових ознак для класифікатора було вперше побудовано алгоритми оцінки синтаксичного та семантичного паралелізму на основі тензорної моделі.
5. Для тензорної моделі мови розроблені алгоритм наповнення багатовимірного тензору опису структур природної мови та потокову архітектуру системи обробки великих текстових корпусів. Тестування системи було проведено на корпусі розміром 100Гб.
6. Для покращення тензорної моделі мови розроблено алгоритм побудови керуючих просторів синтаксичних структур, доведено його коректність та обчислено складність в термінах швидкодії та пам'яті.
7. В результаті реалізації даних алгоритмів було підвищено точність системи знаходження кореферентних зв'язків на 3.5%.

Теоретичне і практичне значення одержаних результатів. Наукове значення роботи полягає в розробці алгоритмів побудови керуючих просторів для природномовних текстів за допомогою конвертації дерев виведення та дерев залежностей, а також у побудові тензорної моделі керуючих просторів на основі виділення закономірностей з багатовимірних частотних словників типових керуючих просторів, що є розвитком математичних моделей представлення семантико-синтаксичних структур текстів.

Практичне значення роботи полягає в покращенні результатів вирішення проблеми побудови відношення кореферентності. Зменшення помилок роботи даної підсистеми автоматично покращує результати наступних прикладних задач:

- автоматичний переклад текстів;
- автореферування;
- природномовні інтерфейси до експертних систем та баз даних
- пошукові системи.

Отримані результати впроваджуються для досліджень в області розробки засобів інтелектуальної обробки текстів природною мовою та для читання

курсів "Штучний інтелект" та "Комп'ютерна лінгвістика" на факультеті кібернетики Київського національного університету імені Тараса Шевченка.

Особистий внесок здобувача. Всі результати дисертаційної роботи отримані автором самостійно, сформульовані у вигляді теорем та алгоритмів та строго доведені з використанням допоміжних лем та тверджень, обґрунтовані з посиланнями на використані джерела.

За результатами дисертації опубліковано вісім робіт у наукових фахових виданнях України [1–5], одна стаття у науковому журналі, внесеному до міжнародних наукометричних баз [1].

У роботах, опублікованих у співавторстві:

– у статті [1] пошукачу належать результати роботи над розробкою моделі природної мови та програмна реалізація алгоритмів побудови розробленої моделі.

– у роботах міжнародних конференцій [6–8] пошукачу належать результати досліджень над розробкою та реалізацією потокової архітектури обробки великих текстових корпусів, уточненням деталей алгоритмів та їх реалізації, проведення експериментів, розроблено методи оцінки якості результатів та проведення тестування згідно з зазначеними методами.

Апробація результатів дисертації. Результати дисертації були представлені на міжкафедральних семінарах факультету кібернетики Київського національного університету імені Тараса Шевченка та Національного університету «Києво-Могилянсько академія» та доповідались на міжнародних конференціях, зокрема на:

1. NLDB’2014, 19-th International Conference on Applications of Natural Language to Information Systems, Natural Language Processing and Information Systems, Montpellier, France, June 18-20, 2014;
2. TSD’2014, 17th International Conference on Text, Speech and Dialogue, Brno, Czech Republic, September 8–12, 2014;
3. PolTAL’2014, 9th International Conference on Natural Language Processing, Warsaw, Poland 17–19 September 2014;

Публікації. За результатами дисертації опубліковано 8 наукових праць, у тому числі 5 – статті у фахових виданнях наукових праць, затверджених МОН України, 1 стаття у журналі, внесеному до міжнародних наукометричних баз, 3 – публікації у матеріалах і працях наукових конференцій.

Структура та обсяг дисертації. Дисертаційна робота складається із вступу, трьох розділів, висновків та списку використаних джерел. Загальний обсяг роботи становить 141 сторінка, основний текст роботи викладено на 95 сторінках, список використаних джерел налічує 102 найменування на 11 сторінках. Текст роботи написаний українською мовою.

ОСНОВНИЙ ЗМІСТ ДИСЕРТАЦІЇ

У **вступі** обґрунтовано актуальність роботи, наведено короткий огляд основних результатів досліджуваної галузі, сформульована мета дисертаційної роботи, проаналізовані основні результати та наведено їх новизну.

У **першому розділі "Аналіз сучасного стану галузі вирішення кореферентностей"** дано системний аналіз існуючих моделей мови та алгоритмів, що використовуються для вирішення кореферентностей.

У **підрозділі 1.1** "Постановка задачі вирішення кореферентностей" визначено місце задачі вирішення кореферентностей в практичних задачах комп'ютерної лінгвістики. Сформульовано означення кореферентного зв'язку та наведені приклади речень з кореферентними сутностями.

У **підрозділі 1.2** "Алгоритми вирішення проблеми анафори" проаналізовано деталі роботи існуючих алгоритмів одного з типів кореферентних зв'язків – анафоричних займенників. Один з перших алгоритмів для вирішення займенникової анафори був розроблений Джеррі Хоббсом[1]. В статті продемонстровано, як за допомогою складної системи семантичного аналізу англомовних текстів можна знаходити пари антецедент-анафора.

Іншим оригінальним алгоритмом вирішення проблеми анафори є алгоритм Шалома Лаппіна та Герберта Лісса[2]. Алгоритм застосовується до формальної моделі тексту, що породжена парсером граматик слотів МакКорда і спирається на міри характерних особливостей, отриманих з синтаксичної структури і динамічної моделі станів.

Один з найбільш розвинених на даний момент алгоритмом вирішення проблеми анафори є алгоритм Міткова[3]. Даний алгоритм є розвитком ідей Лапіна і Ліса. В роботі вводяться нові критерії оцінки: синтаксичний паралелізм, повторюваність кандидата, схоже положення, підмет, доповнення, часте згадування. Також з'являються штрафні критерії, наприклад у випадку не визначеної граматичної ролі слова.

У **підрозділі 1.3** "Алгоритм вирішення кореферентностей" досліджено алгоритми знаходження кореферентностей довільного типу. Один з перших побудованих алгоритмів[4] для знаходження кореферентностей в нерозмічених текстах без обмежень тематики був розроблений в 2001 році. Метод заснований на машинному навчанні. Для кожної пари слів та словосполучень (i,j), що

---

[1] Hobbs J. R. Resolving Pronoun References // Lingua. – 1978. – Vol. 44. – P. 311-338

[2] Shalom L. An Algorithm for Pronominal Anaphora Resolution / Shalom Lappin, Herbert J. Leass // Computational Linguistics. – 1994. – Volume 20 Issue 4. – P. 535-561

[3] Mitkov R. Anaphora resolution (Studies in Language and Linguistics) / Ruslan Mitkov. – Routledge, 2014. – 240 p.

[4] Harabagiu S. M. Text and knowledge mining for coreference resolution / S. M. Harabagiu, R. C. Bunescu, S. J. Maiorano // Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies. – 2001. – P. 1-8

перевіряються на кореферентність, розглядаються 12 простих евристичних критеріїв.

В роботі[1] представлено нове архітектурне рішення розв'язку проблеми кореферентність в вигляді поєднання декількох "решіт". Кожне решето реалізовує деякий алгоритм оцінки пари слів чи словосполучень на кореферентність, працює незалежно, та може повертати одне з трьох значень : сутності кореферентні, не кореферентні чи "не знаю".

Робота Лі[2] є розширенням розробленої в попередній дослідженій системі архітектури заснованої на решетах. В ній вводяться додаткові реалізації решіт, а також алгоритм визначення кандидатів слів та словосполучень на кореферентність. Ця робота продемонструвала найкращі результати на тестовій вибірці конференції CoNLL-2011[3].

У **підрозділі 1.4** "Формальні моделі мови" проаналізовано існуючі моделі мови, що можуть бути використані для розв'язку кореферентностей.

У **пункті 1.4.1** досліджено латентний семантичний аналіз (ЛСА)[4]. ЛСА – це метод обробки інформації природною мовою, що дозволяє проаналізувати взаємозв'язок між колекцією документів і термінами, які в них зустрічаються. Алгоритм зіставляє деякі фактори (теми) всім документам і термам. При класифікації чи кластеризації документів цей метод використовується для вилучення контекстно-залежних значень лексичних одиниць за допомогою статистичної обробки великих корпусів текстів.

Недоліком ЛСА є його обмеженість обробкою двовимірних матриць, а значить виділяються тільки бінарні зв'язки. В природномовних текстах зв'язки в загальному випадку можуть бути складнішими, і охоплювати одразу групу слів. Для вирішення цієї проблеми, було використано невід'ємну факторизацію тензорів.

У **пункті 1.4.2** проаналізовано невід'ємну факторизацію тензорів[5] як розвиток ідеї ЛСА.

Означення. Нехай $I_1, I_2, \ldots, I_N \in \mathbb{N}$ – розміри тензора по кожному з напрямків. В роботі тензор $Y \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$ порядку N визначається як N-вимірний масив з елементами $y_{i_1, i_2, \ldots, i_N}, i_n \in \{1,2,\ldots,I_N\}, 1 \leq n \leq N$.

---

[1] Jonnalagadda S. R. Coreference analysis in clinical notes: a multi-pass sieve with alternate anaphora resolution modules / S. R. Jonnalagadda [and others] // Journal of the American Medical Informatics Association. – 2012. – Volume 19. – P. 867-874

[2] A Multi-Pass Sieve for Coreference Resolution / H. Lee [and others] // Proceedings of the Conference on Empirical Methods in Natural Language Processing. – 2010. – P. 492-501

[3] CoNLL-2011 Shared Task: Modeling Unrestricted Coreference in OntoNotes / S. Pradhan [and others] // Proceedings of the Fifteenth Conference on Computational Natural Language Learning. – 2011. – P. 1-27

[4] Landauer T. Introduction to Latent Semantic Analysis / Thomas K Landauer, Peter W. Foltz, Darrell Laham // Discourse Processes. – 1998. – Volume 25(2-3). – P. 259-285

[5] Tim Van de Cruys A Non-negative Tensor Factorization Model for Selectional Preference Induction // Journal of Natural Language Engineering. – 2010. – Volume 16(4). – P. 517-437

Зовнішній добуток тензорів $Y \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$ та $X \in \mathbb{R}^{J_1 \times J_2 \times \ldots \times J_M}$ визначається за допомогою формули:

$$Z = Y \circ X \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N \times J_1 \times J_2 \times \ldots \times J_M},$$

де $\circ$ – символ операції зовнішнього добутку.

По-елементно операція зовнішнього добутку визначається як:

$$z_{i_1,i_2,\ldots,i_n,j_1,j_2,\ldots,j_m} = y_{i_1,i_2,\ldots,i_n} x_{j_1,j_2,\ldots,j_m},$$

де $z_{i_1,i_2,\ldots,i_n,j_1,j_2,\ldots,j_m}$, $y_{i_1,i_2,\ldots,i_n}$ та $x_{j_1,j_2,\ldots,j_m}$ – елементи тензорів Z, Y, Z відповідно

Особливими випадками є зовнішні добутки двох векторів $a \in \mathbb{R}^I$ та $b \in \mathbb{R}^J$:

$$A = a \circ b = ab^T \in \mathbb{R}^{I \times J}$$

та трьох векторів $a \in \mathbb{R}^I$, $b \in \mathbb{R}^J$ та $C \in \mathbb{R}^Q$:

$$A = a \circ b \circ c \in \mathbb{R}^{I \times J \times Q},$$

де $z_{ijq} = a_i b_j c_q$

Постановка задачі факторизації тензорів. Для даного тензора $Y \in \mathbb{R}^{I \times T \times Q}$ і додатнього індексу J знайти три матриці, що називаються матриці навантаження або фактори, $A = [a_1, a_2, \ldots, a_J] \in \mathbb{R}^{I \times J}$, $B = [b_1, b, \ldots, b_J] \in \mathbb{R}^{T \times J}$, $C = [c_1, c, \ldots, c_j] \in \mathbb{R}^{Q \times J}$, що задовільняють наступним умовам:

$\|E\| \to min$

$$Y = \sum_{j=1}^{J} a_j \circ b_j \circ c_j + E \qquad (*)$$

Умова (*) в поелементній формі:

$$y_{itq} = \sum_{j=1}^{J} a_{ij} b_{tj} c_{qj} + e_{itq}$$

Індекс J в даній задачі називають кількістю фактор-множин.

У **другому розділі "Побудова тензорної моделі керуючих просторів природномовних речень"** розроблено алгоритми побудови формальних моделей мови для заданого речення та встановлено їх обчислювальну складність.

У **підрозділі 2.1** "Аналіз формальної моделі керуючих просторів природномовних речень" проаналізовано зазначену формальну модель синтаксичного представлення[1]. На відміну від суто лінгвістичного підходу, речення в цій моделі розглядається як деякий динамічний обчислювальний рекурсивний процес, що розвивається в керуючому просторі, що пов'язує синтаксично згруповані частини пропозиції інформаційними каналами.

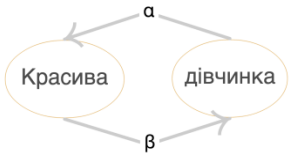
Якщо два об'єкти *A* і *B* вступають у відношення *C*, то ми виділяємо об'єкт (припустимо *A*), що викликає (ініціює, породжує) це відношення, і об'єкт, на який передається це відношення. Таким чином, виділяємо два види спрямованих зв'язків: від об'єкта-генератора відношення до самого відношення і від відношення до підлеглого об'єкту. Перший вид зв'язку називаємо *α*-

[1] Анисимов А. В. Управляющее пространство синт
анализ. – 1990. – №1. – С. 11-17

зв'язком (зв'язок генерування), другий – *β*-зв'язком (зв'язок поширення). Дієслова визначають відносини між об'єктами, тому в стандартній схемі простого речення: “іменник – дієслово – іменник” α-зв'язок направлений від першого іменника до дієслова і β-зв'язок направлений від дієслова до іменника-визначення.

Розглянемо приклад. Дівчинка збирає квіти. Об'єкт “дівчинка” генерує відношення збирає і направляє його на об'єкт “квіти”. Тому *α-β*-структура цього речення має вигляд Рис. 1.а). Розглянемо фразу: Красива дівчинка. Тут об'єкт дівчинка генерує унарне ставлення красива і передає це відношення собі ж (Рис. 1.б). Виникає кільцевий зв'язок, що характеризує зміст словосполучення.

а)



б)

Рис. 1. Керуючий простори фрагментів речень а) “Дівчинка збирає квіти” та б) “Красива дівчинка”.

У **підрозділі 2.2** “Постановка задачі побудови керуючого простору” наведено вхідні та вихідні дані алгоритму та спроектовано рекурсивні структури даних для представлення керуючого простору.

Вхідними даними алгоритму є не розмічений текст природньої мови.

Вихідними даними алгоритму є керуючий простір природномовних текстів в деякому представленні. Для повного опису вихідних даних алгоритму необхідно дати формальний опис цієї структури даних. Проведемо декомпозицію КП та формалізуємо його. Керуючий простір складається з елементів та зв’язків між ними.

Зв’язки бувають 2 типів – альфа та бета. Окрім цього виділимо типовий для КП кільцевий альфа-бета зв’язок в окремий тип зв’язку.

Зв’язок об’єднує 2 елементи утворюючи новий елемент керуючого простору. Елементи утворені таким чином будемо називати складеними. Інформація про тип зв’язку, а також посилання на елементи, що зв’язуються, будуть зберігатися в самих елементах. Крім цього є елементи КП, що вказують на одне слово вхідного речення. Такі елементи будемо називати базовими.

За допомогою розробленої структури даних (Рис. 2.) ми представили керуючий простір синтаксичних структур (Рис. 3. а) в вигляді бінарного дерева (Рис. 3. б).

Теорема 1. Для будь-якого керуючого простору синтаксичних структур природномовних речень існує єдине представлення його в вигляді бінарного дерева

Теорема 2. Для будь-якого представлення КП в вигляді бінарного дерева існує єдиний КП в класичному представленні.

Дані теореми доводять існування бієктивного відображення між двома різними представленнями керуючих просторів.

Рис. 2. UML діаграма структури даних керуючого простору

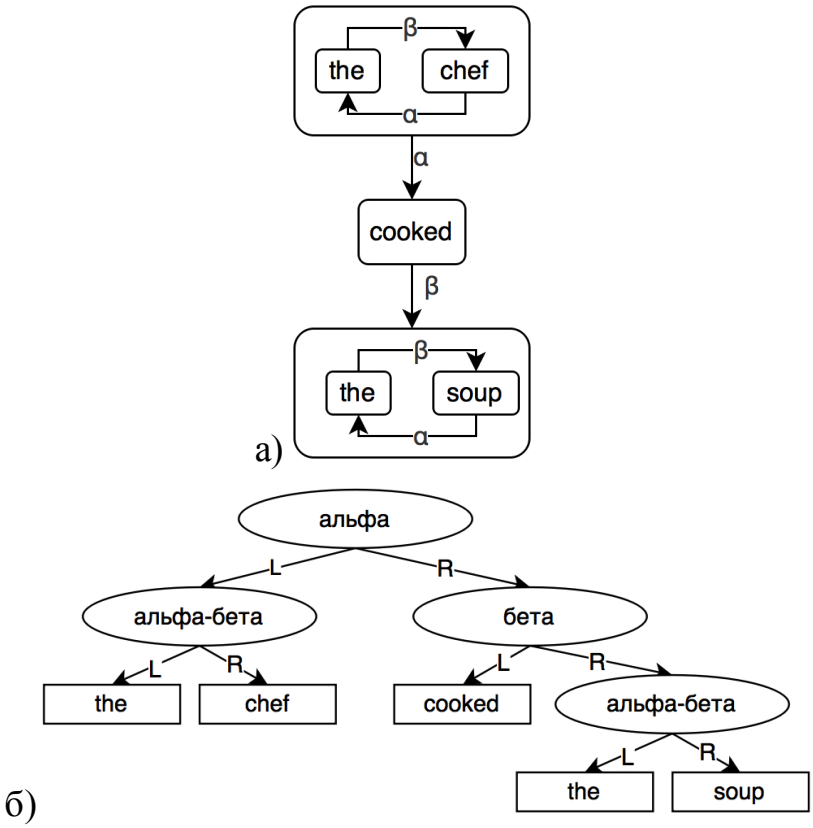


Рис. 3. а) Класичне представлення КП синтаксичних структур. б) представлення КП в формі бінарного дерева

У **підрозділі 2.3** “Попередній синтаксичний аналіз” проведено системний аналіз кожного етапу синтаксичного аналізу текстів з погляду на їх використання при побудові керуючих просторів.

У **пункті 2.3.1** досліджено методи розбиття тексту на речення, проаналізовано типові помилки таких алгоритмів та обгрунтовано методи їх усунення.

У **пункті 2.3.2** проаналізовано проблеми формалізації обсягу лексеми природномовного тексту та досліджено принцип роботи алгоритмів розбиття речення на лексеми.

У **пункті 2.3.3** було сформульовано постановку задачі морфологічного аналізу, досліджено формат вихідних даних системи морфологічного аналізу стенфордського парсера та розглянуто приклади його роботи.

У **пункті 2.3.4** досліджено задачу виділення іменованих сутностей, їх типи, та приклад роботи стенфордського парсера та типові помилки.

У **пункті 2.3.5** розроблено алгоритм виділення іменованих сутностей, що заснований на аналізі сторінок Вікіпедії, та представлено деталі його роботи. Особливу увагу було приділено інтеграції алгоритму в стенфордський парсер.

У **пункті 2.3.6** наведено опис дерев виводу Хомського та проаналізовано приклад дерева виводу природномовного речення англійською мовою побудованого стенфордським парсером.

У **пункті 2.3.7** проаналізовано граматику залежностей, типи дерев залежностей та їх приклади.

У **підрозділі 2.4** “Алгоритм побудови шестивимірного тензора для задачі пошуку прихованих семантичних зв’язків в корпусах природномовних текстів” розроблено нову модель мови, обгрунтовано необхідність її побудови та побудовано алгоритм та структури даних.

У **пункті 2.4.1** обгрунтовано причини збільшення розмірності простору. Тензори розмірності три були введені через неповноту моделей заснованих на тензорах розмірності два.

Проаналізуємо наступний приклад : “радіо грає на піаніно”. Оскільки пари слів речення “радіо грає” та “грає на піаніно” є коректними, то і вся фраза буде вважатися коректною при аналізі за допомогою ЛСА. Розглянемо тензор розмірності три. Його вісі – підмети, присудки та додатки. Трійка “радіо грає на піаніно” буде входити в нього з значенням нуль, оскільки в природномовних текстах такі фрази можуть зустрічатися тільки в казках, а їх ми не будемо брати до уваги.

У **пункті 2.4.2** розроблено рекурсивний алгоритм обчислення елементів тензора з дерев виводу (ДВ) та дерев залежностей (ДЗ). Алгоритм виділяє всі можливі типи зв’язків, з яких на наступному етапі обираються лише шість найголовніших для збереження в тензорі.

У **пункті 2.4.3** було проаналізовано вимоги до способу збереження багатовимірного тензору. Ця задача має важливе значення для аналізу великих

текстових корпусів (Big data). Швидкий доступ до елементів тензору забезпечується за допомогою хеш-таблиць.

У **пункті 2.4.4** було описано використання факторизації природномовних тензорів, обґрунтовано необхідність цього кроку. Для невеликого прикладу проведено всі необхідні обчислення.

У **підрозділі 2.5** "Алгоритм побудови керуючого простору природномовних речень" детально описано принцип роботи розробленого алгоритму. Входом алгоритму є дерево залежностей, дерево виводу та інші результати попереднього синтаксичного аналізу. Виходом алгоритму є керуючі простори природномовних структур.

У **пункті 2.5.1** побудовано алгоритм конвертація типу зв'язку дерева залежностей в тип зв'язку керуючого простору. Цей алгоритм вирішує часткову задачу – визначення типу зв'язку КП між існуючими КП, пов'язаних між собою зв'язком ДЗ.

Головна ідея алгоритму полягає в припущенні, що кожному з типу зв'язків дерева залежностей відповідає один з типів зв'язку керуючого простору – альфа-, бета-, та кільцевий альфа-бета-зв'язок певного напрямку. Конвертація типу відбувається за допомогою звернення до попередньо побудованої хеш-таблиці словника. За рахунок вдалого вибору хеш-функції та завантаження словника типів повністю в пам'ять на етапі попередньої обробки даних визначення типу зв'язку КП відбувається за константний час.

Лема 1. Алгоритм конвертації працює за час *O(1)* та потребує *O(N)* додаткової пам'яті, де *N* – кількість типів зв'язків Дерева залежностей.

У **пункті 2.5.2** наведено та обгрунтовано алгоритм створення елементу керуючого простору для двох слів. Нехай дано 2 слова, їх POSTag'и, тип зв'язку ДЗ між ними та тег найменшого піддерева дерева виводу, що їх містить. Тоді тегом синтаксичного значення КП буде тег найменшого піддерева виводу. Тип зв'язку визначається за допомогою алгоритму конвертації типу зв'язку.

У **пункті 2.5.3** розроблено алгоритм створення складного елементу керуючого простору для піддерева дерева виводу. Нехай для деякого вузла дерева виведення дано його прямі нащадки – піддерева з відповідними вже побудованими керуючими просторами. Рівні ієрархії в керуючому просторі для них будемо визначати пріоритетом тегів дерева залежностей. Найвищий пріоритет мають конкретні та специфічні типи зв'язку, що впливають лише на маленькі групи слів (наприклад зв'язок типу прикметник-іменник).

Розглянемо оптимізований алгоритм об'єднання:

```
function об'єднатиДекількаКПВОдин
    (Базовий_елемент кп[]) : Базовий_елемент
begin
    while sizeof(кп) > 1
    begin
        найбільшийПріоритет ← 0
        параЗНайвищимПріорітетом ← -1
```

```
        n ← sizeof(кп)
        for i ← n step -1 until 2
        begin
            if існує зв'язок зв в граматичних відносинах
            між словами кп[i] і кп[i-1] and зв.пріоритет
            > найбільшийПріоритет then
            begin
                найбільшийПріоритет ← зв.пріоритет
                параЗНайвищимПріорітетом ← i
            end
         end
        if параЗНайвищимПріорітетом != -1 then
        begin
            i ← параЗНайвищимПріоритетом
            замінити кп[i] та кп[i-1] простором
            об'єднати2КП(кп[i], кп[i-1], тип граматичного
            зв'язку)
        end
    end
    return кп[0]
end
```

Лема 2. Оптимізований варіант алгоритму об'єднання працює за час $O(N^2)$, та потребує $O(M)$ додаткової пам'яті, де N – кількість КП, що необхідно об'єднати, M – кількість типів зв'язків КП.

У **пункті 2.5.4** розроблено та досліджено узагальнюючий алгоритм побудови КП, що полягає рекурсивному обході дерева виводу та використанні описаних вище допоміжних алгоритмів.

```
function побудуватиКП(Дерево р) : Базовий_елемент
begin
    if р – листок
    then return Базовий_елемент(синтаксичний тег
    отриманим з морфологічної розмітки, слово-термінал
    для цього листку)
    else
    begin
        сини ← список синів вершини р
        for син ← сини
            кп ← union(кп,  {побудуватиКП(син)})
        return
             об'єднатиДекількаКПВОдин(керуючіПростори)
    end
end
```

Теорема 3 (про обчислювальну складність). Алгоритм побудови КП працює за час $O(N*K)$ та потребує $O(M+K*log_K(N))$ пам'яті, при умові, що кожен вузол має K синів, N – сумарна кількість терміналів в дереві, M – кількість типів дерева залежностей. $K>1$

При доведення часової складності алгоритму було визначено, що кількість викликів алгоритму об'єднання з часовою складністю $O(K^2)$ має порядок $O(N/K)$.

Теорема 4 (повна коректність). Алгоритм будує керуючий простір за скінченну кількість кроків, що відповідає вхідним Дереву виводу та Дереву залежностей.

У **пункті 2.5.5** обґрунтовано метод покращення створеного КП за допомогою використання інформації про виділені іменовані сутності. Результат оформлено як алгоритм попередньої модифікації дерева виводу перед викликом алгоритму побудови КП.

У **підрозділі 2.6** "Виділення типових елементів природномовних речень" спроектовано систему для обробки великих текстових корпусів.

У **пункті 2.6.1** розроблено ER-модель бази даних для збереження тензора. Для створеної моделі було доведено відповідність нормальній формі Бойса-Кода.

У **пункті 2.6.2** було розроблено конвеєрну архітектуру обробки великих текстових корпусів. Побудовано наступні фільтри конвеєру: читання файлу, розархівування потоку даних, розбір XML формату за допомогою SAX та інтерфейс процесора сторінок Вікіпедії.

У **підрозділі 2.7** "Профілювання та оцінка результатів" було оцінено якість та швидкодію розробленої підсистеми. Оцінку якості роботи алгоритму представлено в термінах точності та покриття. Результати наведено окремо по кожному з типів зв'язків КП та в цілому. Проведено експериментальне встановлення швидкодії реалізації алгоритму за допомогою засобів профілювання додатків. В середньому алгоритм побудови керуючого простору вимагає менше ніж на порядок часу ніж попередній синтаксичний аналіз. Обробка одного рівня Дерева виведення займає близько 400 мілісекунд на одному ядрі процесора з тактовою частотою 2.5 ГГц.

У **третьому розділі "Алгоритми визначення кореферентних зв'язків за допомогою статистичної інформації типових структур керуючих просторів"** було використано результати попереднього розділу для вирішення практичної задачі комп'ютерної лінгвістики. Для визначення кореферентностей використано архітектуру засновану на ситах (фільтрах), до якої було додано декілька нових решіт, що реалізують класифікатор методу опорних векторів.

У **підрозділі 3.1** "Пошук сутностей" проаналізовано принципи роботи алгоритму виділення сутностей в тексті за допомого інформації представленою деревом виводу. З дерева виводу обираються всі іменникові фрази (мають тег NP та NNP). Також до переліку знайдених сутностей додаються іменовані сутності (власні назви). В результаті отримуємо список кандидатів в сутності

для знаходження кореферентних зв'язків. Наступним кроком список фільтрується згідно визначених правил. Наприклад видаляються сутності, що входять в інші сутності, тому що вони посилаються на один і той же об'єкт, але довший варіант цієї сутності містить більше додаткової інформації.

У **підрозділі 3.2** "Зведення задачі знаходження кореферентностей до тернарного класифікатора пари сутностей" досліджено метод зведення задачі кластеризації кореферентностей до задачі класифікації. Спочатку список виділених сутностей сортується за номером речення. Всередині одного речення сутності сортуються за відстанню від кореня дерева виводу. В кожному новоутвореному кластері сутностей ми будемо порівнювати тільки перші з них. Спочатку буде застосовано перше решето до всіх пар сутностей, потім друге і так далі.

У **підрозділі 3.3** "Порівняння піддерев" досліджені принципи синтаксичного та семантичного паралелізму та розроблено алгоритм реалізації цих принципів за допомогою методу оцінки подібності дерев керуючих просторів синтаксичних структур для пари сутностей. Побудований алгоритм рекурсивно обходить одночасно два КП, аналізуючи типи зв'язків та семантичну близькість слів за допомогою порівняння фактор-множин розкладених частотних словників типових КП.

Теорема 5. Алгоритм оцінки семантико-синтаксичного паралелізму працює за час $O(min(N_1, N_2)K)$ та використовує $O(log_2 min(N_1, N_2)$ пам'яті, де N_1 та N_2 – кількість елементів в КП, що аналізуються, K – кількість фактор-множин, обрана при факторизації тензора.

У **підрозділі 3.4** "Оцінка наявності кореферентного зв'язку за допомогою виявлення семантичного паралелізму керуючих просторів" розроблено алгоритм виявлення кореферентних зв'язків.

Вхід алгоритму : дві сутності "а" та "б" та відповідні їм керуючі простори синтаксичних структур "А" та "Б".

Вихід : оцінка сумісності.

Основні кроки алгоритму:

Крок 1. Підставимо слово чи словосполучення сутності "а" в керуючий простір "Б" замість слова чи словосполучення сутності "б". Отримаємо фрагмент КП "Б" з сутністю "а".

Крок 2. Зробимо запит до тензору відповідних зв'язків для даного елементу.

Крок 3. Результатом запиту буде перше число, що характеризує оцінку семантичної сумісності першої сутності в контексті другої сутності.

Теорема 6. Алгоритм оцінки семантичного паралелізму працює за час $O(K)$, де K – кількість фактор-множин, обрана при факторизації тензора.

У **підрозділі 3.5** "Визначення слів-індикаторів як фактор-множин розкладених тензорів" розвивається ідея використання слів індикаторів для вирішення кореферентностей. В розробленому алгоритмі список слів індикаторів будується автоматизовано, на відміну від традиційного підходу, де список слів будується повністю вручну. Крім того було зменшено

обчислювальну складність роботи порівняно з традиційним підходом, що працює за час *O(N)*, де *N* – кількість слів-індикаторів, наступним чином:

Теорема 7. Алгоритм оцінки наявності кореферентного зв'язку за допомогою фактор-множин працює за час *O(K)*, де *K* – кількість фактор-множин використаний в невід'ємній факторизації.

У **підрозділі 3.6** "Особливості тензорної факторизації розгорнутих та кільцевих альфа-бета зв'язків" проаналізовано алгоритм поблокового координатного спуску, що дозволяє провести невід'ємну факторизацію великих текстових корпусів частинами.

У **підрозділі 3.7** "Модифікація алгоритму опорних векторів для великих об'ємів негативних прикладів" було проаналізовано метод опорних векторів та тренувальну вибірку та модифіковано алгоритм навчання, виходячи з їх особливостей. Метод опорних векторів – це набір схожих алгоритмів виду «навчання із вчителем»[1]. Основна ідея методу опорних векторів – перевід вихідних векторів у простір більш високої розмірності та пошук роздільної гіперплощини з максимальним проміжком у цьому просторі

В використаній навчальній вибірці є близько 20 тисяч кореферентних пар сутностей та більше мільйона не кореферентних. Головна ідея модифікації алгоритму полягає в генеруванні тренувальної вибірки під час процесу навчання. В цільову тренувальну вибірку додаються тільки ті негативні приклади, результат класифікації яких не збігається з еталонним.

У **підрозділі 3.8** "Простір ознак машинного навчання" побудовано простір з 28 різноманітних ознак пари сутностей для побудови класифікатора методом опорних векторів. Отримуємо точку в 28-вимірному просторі ознак, що належить до відповідного класу. Метод опорних векторів будує гіперплощину, що розділяє точки класів кореферентних та не кореферентних сутностей.

У **підрозділі 3.9** "Оцінка результатів роботи алгоритму пошуку кореферентностей" детально описано процедуру тестування CoNLL'2011. Точність роботи алгоритмів пошуку кореферентностей визначається за допомогою експерименту.

Процедура визначення точності на тестовій вибірці полягає в оцінці правильності визначених кореферентних кластерів сутностей. Оцінка точності роботи системи за мірою MUC обчислюється за допомогою наступної формули:

$$MUC = \frac{\sum_{i=0}^{n'} |C'_i \cap C_{j(i)}| - 1}{\sum_{i=0}^{n} |C_i| - 1}, j(i) = argmax(|C'_i \cap C_j|), j = 0,1,\dots,n,$$

де C_i – множина сутностей в і-тому кластері ручної розмітки,

C'_i – множина сутностей в і-тому кластері отриманий за допомогою досліджуваної системи,

[1] Burges C.J.C. A Tutorial on Support Vector Machines for Pattern Recognition // Data Mining and Knowledge Discovery. – 1998. – Volume 2. – P. 121-167

j(i) – номер кластера досліджуваної системи, що має найбільший перетин з кластером i з ручної розмітки. Вважається, що кластер з найбільшим перетином є відповідником кластера з ручної розмітки.

Міра MUC для побудованої в роботі системи дорівнює 64.45%, що майже на 3.5% більше ніж попередній найкращий результат Stanford Deterministic Coreference Resolution System, що дорівнює 61.03%.

ВИСНОВКИ

Основним результатом дисертації є розробка та математичне обґрунтування нових алгоритмів ідентифікації та аналізу кореферентних зв'язків у природномовних текстах, що має істотне значення для розв'язання фундаментальної задачі комп'ютерної лінгвістики - семантичного аналізу текстів. Для цього було застосовано тензорну модель природної мови, керуючі простори синтаксичних структур речень та методи машинного навчання. При виконанні роботи одержано такі наукові результати:

1. В методі опорних векторів удосконалено алгоритм навчання для класифікації кореферентних сутностей. Це дало змогу одержати більш точні результати класифікації для типової задачі знаходження кореферентностей, коли кількість не кореферентних пар слів на декілька порядків перевищує кількість кореферентних пар.
2. Для підвищення точності класифікації було розроблено розширений простір ознак із додаванням семантико-синтаксичних властивостей. Для обчислення параметрів кореферентних пар в розширеному просторі ознак було вперше побудовано алгоритми оцінки синтаксичного та семантичного паралелізму на основі тензорної моделі.
3. Для тензорної моделі мови розроблено алгоритм побудови багатовимірного масиву опису структур речень та потокову архітектуру системи обробки великих текстів. Тестування системи було проведено на наборі текстів сумарним розміром 100Гб.
4. Розроблено новий алгоритм побудови керуючих просторів синтаксичних структур речень, який дозволив отримати зручне та стисле представлення моделі, зменшити розмірність тензора, отримати більш надійний та стійкий опис семантико-синтаксичних зв'язків між словами. Доведено коректність та обчислено складність алгоритму в термінах швидкодії та пам'яті.
5. Для тестування розробленої системи використовувалася введена конференцією CoNLL-2011 вибірка, яка є стандартом для аналізу роботи систем кореферентних зв'язків. В результаті інтеграції розроблених алгоритмів в одну з найкращих систем визначення кореферентних зв'язків Stanford Deterministic Coreference Resolution вдалось покращити за запропонованою на конференції MUC-6 мірою точність визначення на вказаній тестовій вибірці з 61.03% до 64.45%.

СПИСОК ОПУБЛІКОВАНИХ ПРАЦЬ ЗА ТЕМОЮ ДИСЕРТАЦІЇ

Статті у фахових виданнях:

1. Вознюк Т. Г. Определение семантических валентностей концептов онтологий с помощью неотрицательной факторизации тензоров больших текстовых корпусов / А. В. Анисимов, А. А. Марченко, Т. Г. Вознюк // Кибернетика и системный анализ, 2014, №3, с. 3-16
2. Вознюк Т. Г. Алгоритм побудови керуючого простору синтаксичних структур природномовних текстів // Вісник Київського національного університету імені Тараса Шевченка, Серія фізико-математичні науки, 2014, №1, с.122-127
3. Вознюк Т. Г., Алгоритм побудови шестивимірного тензора для задачі пошуку прихованих семантичних зв'язків в корпусах природномовних // Проблеми програмування, 2014, №2-3, с. 273-278
4. Вознюк Т. Г., Застосування керуючого простору синтаксичних структур природномовних текстів для вирішення проблеми анафори // Вісник Київського національного університету імені Тараса Шевченка, Серія фізико-математичні науки, 2014, №2, с. 101-105
5. Вознюк Т. Г., Побудова класифікатора для вирішення займенникової анафори на основі тензорної моделі // Вісник Київського національного університету імені Тараса Шевченка, Серія фізико-математичні науки, 2015, №1, с. 66-70

Матеріали та праці конференцій:

6. Anisimov A. V. Semantic and Syntactic Model of Natural Language Based on Tensor Factorization / A.V. Anisimov, T. G. Vozniuk, V. Taranukha // NLDB-2014, Lecture Notes in Computer Science. – Volume 8455. – Springer Verlag. – 2014 .– PP. 51–54.
7. Anisimov A. V. Development of a semantic and syntactic model of natural language by means of non-negative matrix and tensor factorization / A.V. Anisimov, T. G. Vozniuk, V. Taranukha // TSD-2014, Lecture Notes in Artificial Intelligence. – Volume 8655. – Springer Verlag. –2014.– PP. 324–335.
8. Anisimov A. V. Semantic and Syntactic Model of Natural Language Based on Non-negative Matrix and Tensor Factorization / A.V. Anisimov, T. G. Vozniuk, V. Taranukha // PolTAL-2014, Lecture Notes in Artificial Intelligence. – Volume 8686. – Springer Verlag. –2014.– PP. 177–184.

АНОТАЦІЯ

Вознюк Т.Г. *Застосування семантико-синтаксичної тензорної моделі природної мови для аналізу кореферентних зв'язків у текстах. – Рукопис.*

Дисертація на здобуття наукового ступеня кандидата фізико-математичних наук за спеціальністю 01.05.01 – теоретичні основи інформатики та кібернетики.

– Київський національний університет імені Тараса Шевченка МОН України. – Київ, 2015.

Основним результатом дисертації є розробка та математичне обґрунтування нових алгоритмів ідентифікації та аналізу кореферентних зв'язків у природномовних текстах, що має істотне значення для розв'язання фундаментальної задачі комп'ютерної лінгвістики - семантичного аналізу текстів. Для цього було застосовано тензорну модель природної мови, керуючі простори синтаксичних структур речень та методи машинного навчання.

В дисертаційній роботі розроблено новий алгоритм побудови керуючих просторів синтаксичних структур речень, який дозволив отримати зручне та стисле представлення моделі, зменшити розмірність тензора, отримати більш надійний та стійкий опис семантико-синтаксичних зв'язків між словами.

В результаті інтеграції розроблених алгоритмів в одну з найкращих систем визначення кореферентних зв'язків Stanford Deterministic Coreference Resolution вдалось покращити за запропонованою на конференції MUC-6 мірою точність визначення на тестовій вибірці конференції CoNLL-2011 з 61.03% до 64.45%.

Ключові слова: кореферентні зв'язки, кореферентність, анафора, тензорні моделі мови, керуючі простори синтаксичних структур, машинне навчання, метод опорних векторів.

АННОТАЦИЯ

Вознюк Т.Г. *Применение семантико-синтаксической тензорной модели естественного языка для анализа кореферентных связей в текстах. Рукопись*.

Диссертация на соискание ученой степени кандидата физико-математических наук по специальности 01.05.01 – теоретические основы информатики и кибернетики. – Киевский национальный университет имени Тараса Шевченко МОН Украины. – Киев, 2015.

Диссертация посвящена улучшению работы систем нахождения кореферентных связей в текстах с помощью тензорной модели естественного языка. При разработке новых алгоритмов решения этой проблемы были детально проанализированные разработки Хоббса, Лаппина и Лиса, Миткова и Толпегина для решения частичной проблемы нахождения кореферентных связей – разрешение проблемы анафоры. Одна из первых работ по решению общей задачи нахождения кореферентностей датируется 2001 годом и принадлежит группе ученых из Сингапура, которые использовали 12-мерное пространство признаков для обучения классификатора.

С помощью потоковой архитектуры обработки данных и модулей предварительного синтаксического анализа был проанализирован большой текстовый корпус объемом 100Гб и построено шестимерный тензор типичных синтаксических структур. Алгоритмы наполнения тензора принимают на вход Деревья вывода и Деревья зависимостей.

Для улучшения построенной тензорной модели были разработаны новые алгоритмы построения управляющих пространств (УП) естественно-языковых предложений, позволяющих сохранить больше информации о семантико-синтаксических связях в тензорах меньшего размера. В рамках работы были разработаны структуры данных УП синтаксических структур и алгоритмы их автоматического построения. Для представления УП было спроектировано рекурсивную структуру данных. Базовый элемент УП представляет подпространство, которое содержит только одно слово. В качестве атрибутов этого элемента задаются тег синтаксического значения, в данном случае совпадающим с тегом морфологического значения слова. Сложенный элемент УП наследуются от базового элемента, поэтому содержит все его атрибуты. Тег синтаксического значения сложенного элемента совпадает с тегом минимального уровня иерархии Дерева вывода, который полностью содержит это УП. Также сложенный элемент УП содержит новые атрибуты : ссылки на 2 элемента УП (базовый или сложенный) и тег типа связи УП. Комбинируя эти два типа элементов можно построить любое валидное УП.

Алгоритм построения УП для удобства восприятия было разбитого на несколько логических этапов, каждый из которых представлен своим алгоритмом. Первый этап – это алгоритм конвертации типа связи Дерева зависимостей в тип связи УП. Наибольшую сложность данного этапа представляет непосредственно создание словаря соответствий типов связей. Этот алгоритм используется на следующем этапе для построения УП одного уровня иерархии Дерева зависимостей, при известных построенных УП для поддеревьев данного уровня. Следующим логическим этапом является алгоритм рекурсивного обхода Дерева вывода. Алгоритм обходит Дерево вывода снизу к верху объединяя уже построенные УП для вершин-сынов.

Для подсчета вектора признаков в машинном обучении использована тензорная модель языка и управляющее пространство синтаксических структур. С помощью выделения типичных структур управляющих пространств был модифицирован алгоритм оценки влияния слов индикаторов на наличие кореферентной связи между пару слов, что позволило строить список слов индикатор полуавтоматически. Идеи семантического и синтаксического параллелизма, которые заключаются в проявлении схожих сематических и синтаксических свойств кореферентными сущностями, впервые были с помощью тензорной модели языка. Это позволило покрыть большее число ситуаций в которых такие алгоритмы выдают корректный результат. Было рассчитано алгоритмическую сложность работы разработанных алгоритмов в нотации Большой О. Результаты оценки сложности были оформлены в виде теорем с доказательствами.

В работе проведен ряд экспериментов по тестированию разработанной системы. Отдельно было протестировано качество разработанных алгоритмов построения управляющих пространств в терминах покрытия и точности по каждому из типов связи отдельно. Тестовая выборка конференции CoNLL-2011

Shared Task: Modeling Unrestricted Coreference in OntoNotes, что использовалась в экспериментах, является стандартом для тестирования систем анализа кореферентных связей. В результате реализации разработанных в работе алгоритмов результаты работы Стэндфордской системы кореферентного анализа, одной из лучших существующих в мире системе (state of the art), были улучшены с 61.03% до 64.45% за метрикой MUC.

Ключевые слова: кореферентные связи, кореферентность, анафора, тензорные модели языка, управляющие пространства синтаксических структур, машинное обучение, метод опорных векторов.

ABSTRACT

Vozniuk T.G. *The use of semantic-syntactic tensor model of natural language for analysis of coreferential relations in texts. Manuscript.*

The thesis for the degree of Candidate of Physical and Mathematical sciences on a speciality 01.05.01 – the theoretical foundations of computer science and cybernetics. – Taras Shevchenko National University of Kyiv, Ministry of Education and Science of Ukraine. – Kiev, 2015.

Thesis is dedicated to improvement of systems for finding coreferential connections in the text using the tensor model of natural language. A large 100Gb text corpus was analyzed and a six-dimensional tensor of typical syntactic structures was built using pipeline data processing architecture and syntactic structures parsers. New efficient algorithms for constructing the control space of syntactic structures was developed to improve the tensor model. This made possible to save more information about the semantic and syntactic relationships in tensors of less dimension.

The developed algorithms were used to solve the problem of finding coreferential connections in natural language texts. It was decided to use multi-sieve approach for Coreference resolution, because it demonstrated best result on well-known sets. The paper describes the new sieves designed for the system, which implements classifier of support vector machine for the 28-dimensional feature space. The tensor model of language and control space syntactic structures were used to calculate the feature vector used in machine learning.

The series of experiments was designed to test the system. The test set of the conference CoNLL-2011 Shared Task: Modeling Unrestricted Coreference in OntoNotes, that was used in the experiments is the standard for testing systems of coreferential analysis. Extending of Stanford coreferent analysis system, one of the best of the world's system (state of the art), by implementation of the algorithm proposed in the work improved results from 61.03% to 64.45% by the MUC metric.

Keywords: coreference resolution, coreference, anaphora, tensor model of language, the control space of syntactic structures, machine learning, support vector machine.