УДК 681.3.062

Вознюк Т.Г.[1], аспірант.

Застосування керуючого простору синтаксичних структур природномовних текстів для вирішення проблеми анафори

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

T.G. Vozniuk[1], Postgraduate Student.

An application of the control space of syntactic structures of natural language sentences for the anaphora resolution

[1] Taras Shevchenko National University of Kyiv, 83000, Kyiv, Glushkova st., 4d,
e-mail: taarraas@gmail.com

В роботі обгрунтовано можливість застосування керуючого простору синтаксичних структур природномовних текстів для вирішення проблеми анафори. Наведено короткий опис проблеми та відомих методів її вирішення. Розглянуто модель керуючих просторів та проілюстровано використання латентних семантичних зв'язків, виділених з частотних словників складових керуючих просторів.

Ключові слова: штучний інтелект, комп'ютерна лінгвістика, керуючий простір, проблема анафори.

This work justifies the possibility of using the control space of syntactic structures of natural language texts for the anaphora resolution. The anaphora resolution approaches by Hobbs, Mitkov, Lappin and Leass were briefly described. The model of the control space was considered. This model consists of alpha- and beta-connections between words and word combinations on the few levels of hierarchy. There was formulated a task as an assessing the likelihood that a pronoun anaphora refers to a word or phrase on the basis of their use in the context of control spaces using The Tensor Model of language. The algorithm describes the transformation of the control spaces of analyzed sentences and requests to The Tensor Model. Overall score is defined as the weighted sum of indicators of simple criteria. There was formulated optimization task for training of weights. An example of using semantic relations extracted from the frequency vocabulary of components of control spaces latent was given. For a simple example score of the correct pair antecedent-anaphora in the meaning of semantic was significantly higher than the incorrect pair. This shows an ability of the proposed model to solve semantic anaphora ambiguities.

Keywords: artificial intelligence, computational linguistics, control space, anaphora resolution

Статтю представив д.ф.-м.н., проф. Анісімов А.В.

Аналіз природномовних текстів є невід'ємною складовою задач автоматичного перекладу текстів, інтелектуального пошуку, реферування та реалізації природномовних інтерфейсів. Для більш повного аналізу природномовних текстів визначають семантичні зв'язки між словами та словосполученнями. Одним з таких зв'язків є зв'язок між анафоричним займенником та антецедентом.

Анафоричні займенники – це займенники, що відсилають до деякого іншого слова чи словосполучення (антецедент) цього тексту, семантичне значення якого вони відображають. Вирішення проблеми анафори полягає в встановленні відповідностей між анафоричними займенниками та антецедентами.

Наприклад:

"Дівчинка побачила хлопчика. Він збирав ягоди."

В данному випадку «Він» - анафоричний займенник. «Дівчинка» і «хлопчика» - антецеденти. Вирішивши проблему анафори для даного речення ми зв'язуємо займенник «Він» з антецедентом «хлопчика», використовуючи морфологічні ознаки роду.

Для вирішення анафори за допомогою алгоритму Лапіна і Ліса[1] використовують оцінки близькості речень анафори і андецедента,

підметниковий наголос, наголос існування, пряме та не пряме доповнення. Остаточна оцінка схожості визначається як зважена сума визначених вище оцінок.

В алгоритмі Хоббса[2] використовується синтаксична структура дерева виведення, яка вважається наперед визначеною. Ідеєю метода є оригінальний алгоритм обходу дерева виводу, в процесі якого визначаються пари анафора-антецедент.

Алгоритм Міткова[3] розвиває ідею Лапіна і Ліса, проте 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.

Кожний з цих методів пропонує свої засоби оцінки та варіанти їх інтеграції. В даній роботі буде запропонована одна з можливих оцінок, що грунтується на латентних семантичних зв'язках керуючого простору синтаксичних структур. Вона дасть змогу скористатися високими результатами моделі керуючих просторів для розв'язання проблеми анафори.

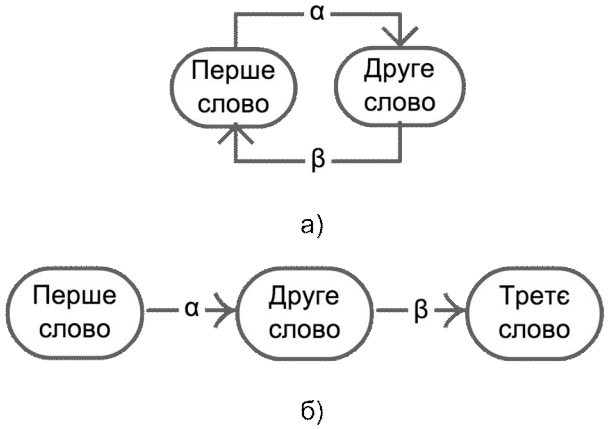
Синтаксична та семантична модель природньої мови, що заснована на статистичній інформації про керуючі простори синтаксичних структур природньомовних текстів

В основі ідеї моделі лежить симбіоз апарату керуючих просторів (КП), латентного семантичного аналізу, методу невід'ємної факторизації матриць та тензорів високих порядків.

Керуючий простір представляє собою комбінацію деяких зв'язків між елементами. В якості елементів КП синтаксичних структур виступають слова та групи слів. Зв'язки визначаються наступним чином: 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 перекладається це відношення. Таким чином виділяємо два види мовних зв'язків: від об'єкта-генератора ставлення до відношення і від ставлення до підлеглого об'єкту. Перший вид зв'язку називаємо альфа-зв'язком (зв'язок генерування), другий - бета-зв'язок (зв'язок поширення)[4].

В [5] запропоновано алгоритм автоматичної побудови керуючого простору для довільного природномовного речення. В [6] описано процес побудови моделі синтаксичних і семантичних зв'язків на основі статистичної інформації виділеної з керуючих просторів великих текстових корпусів. Ця модель лежить в основі запропонованого в статті методу.

Модель була навчена на великих текстових корпусах об'ємом порядку 100 Гб, а тому містить відомості вживання слів в різних значеннях різноманітних тематик.



Мал. 1. а) Графічне зображення двійки слів моделі – кільцевого альфа-бета зв'язку б) Графічне зображення трійки слів моделі

В КП природномовних текстів можна виділити двійки та трійки слів. Двійка в розумінні моделі – це пара слів, з'єднаних кільцевим альфа-бета зв'язком. Трійка – це трійка слів, зв'язаних направленими альфа- та бета- зв'язками. Порядок з'єднання вказаний на Мал. 1. Так як в цій моделі речення представлені комбінацією альфа- та бета-зв'язків, то кожне з них можна розкласти в двійки та трійки, при цьому кожне слово вихідного речення ввійде у принаймні одну двійку чи трійку.

Латентний семантичний аналіз представлений двома простими запитами до моделі: взнати оцінку вірогідності вживання двійки та трійки в природньомовному тексті. Комбінуючи види запитів та повторюючи ієархічну структуру керуючого простору, ми отримуємо можливість оцінювати більш складні структури природномовних речень.

Алгоритм критерію оцінювання пари антецедент-анафора

Нехай дано слово чи словосполучення антецеденту w_i та речення в яке воно входить s_i. Також маємо анафоричний займенник w' та речення в яке він входить s'.

Задача: дати оцінку ймовірності того, що анафоричний займенник w' посилається на слово чи словосполучення w_i виходячи з їх вживанн в контексті речень s_i та s' відповідно.

Алгоритм:

Побудувати керуючі простори cs_i та cs' для речень s_i та s' за допомогою алгоритму запропонованого в [2].

Визначити двійку two_i та трійку $three_i$, якщо вони існують, такі що :

а) two_i in cs_i, $three_i$ in cs_i

б) w_i входить як компонент в two_i і $three_i$

Визначити двійку two' та трійку three', якщо вони існують, такі що :

а) two' in cs', three' in cs'

б) w' входить як компонент в two' і three'

Визначити оцінки схожості контекстів в термінах керуючих просторів :

$$score_{two} = c_1*\{two_i.first = two'.first\} + c_2*\{two_i.second = two'.second\} + c_3*\{two_i.type = two'.type\}$$

$$score_{three} = c_4*\{three_i.first = three'.first\} + c_5*\{three_i.second = three'.second\} + c_6*\{three_i.third = three'.third\}$$

Визначити оцінку вживаності фрази за допомогою описаної вище моделі природньої мови. Для цього необхідно сформувати нову двійку two'' та трійку three'' замінивши в них входження анафоричного займенника w' на слово w_i, та визначити оцінки $score\_two^*$ та $score\_three^*$ по розробленій моделі.

Знайти інтегральну оцінку як лінійну комбінацію наведених вище оцінок:

$$score = c_7*score_{two} + c_8*score_{three} + c_9*score_{two}^* + c_{10}*score_{three}^*$$

Коефіцієнти c_i, $\overline{i = 1,10}$ можна знайти поставивши задачу оптимізації покриття (recall) при обмеженому показнику точності (precision) деяким мінімальним значенням $precision_{min}$:

$$\bar{c} = \arg\min_{precision(корпус,\bar{c}) > precision_{min}} recall(корпус,\bar{c})$$

Таку задачу можна вирішити за допомогою алгоритму імітації відпалу[7].

Описаний метод оцінки може застосовуватися як єдина метрика, чи в комбінації з іншими евристиками запропонованими в роботах Міткова, Лаппіна та Ліса, тощо. Комбінація запропонованого алгоритму з іншими евристиками, при умові тренування вагових коефіцієнтів кожної з оцінок на великих текстових корпусах, повинні дати кращі результати вирішення проблеми анафори.

Розглянемо принцип роботи запропонованого алгоритму на наступному прикладі:

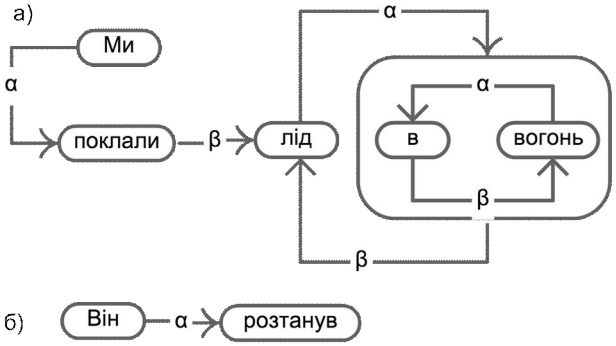
Ми поклали лід в вогонь. Він розтанув.

Знайдемо оцінку пари вогонь-розтанув.

Вхідні дані :

w_1 = "вогонь". s_1 = "Ми поклали лід в вогонь."

w' = "він". s' = "Він розтанув."



Мал. 2. а) КП речення «Ми поклали лід в вогонь» б) КП речення «Він розтанув»

Керуючі простори для речень s_1 та s' представлені на Мал. 2.

В реченні s_1 слово «вогонь» входить тільки в двійку two_1 = (в, вогонь)

В реченні s' «він» входить в трійку three' = (він, розтанув, -).

Так як в двійку слово входить тільки в першому реченні, то $score_{two}$ = 0.

Аналогічно $score_{three}$ = 0.

Так як двійка в другому реченні відсутня, то $score_{two}^*$ = 0.

Оцінимо значення $score_{three}^*$. Для цього необхідно слово «він» замінити на слово «вогонь» в трійці three' = (він, розтанув, -). В результаті отримаємо трійку three'' = (вогонь, розтанув, -). Оцінка цієї трійки згідно моделі керуючих просторів буде близькою до нуля, оскільки «вогонь» не може входити в зв'язок генерування з словом «розтанув»

Отже, сумарна оцінка пари вогонь-розтанув буде досить малою, що й очікувалося.

Аналогічно розглянемо пару лід-розтанув. Метрики $score_{two}$, $score_{three}$, $score_{two}^*$ будуть дорівнювати нулю, оскільки тип підпросторів в які входять слова не змінився, а тому їм так само не буде відповідників. Ситуація з $score_{three}^*$, буде інакшою, оскільки оцінка по моделі трійки (лід,

розтанув, -) має суттєво відрізнятися від нуля, а тому й інтегральна оцінка буде значно вищою ніж для пари вогонь-розтанув. Система в такому випадку зробить правильне рішення про посилання анафоричного займенника «він» на антецедент «лід».

Висновки

У роботі був запропонований алгоритм вирішення проблеми анафори за допомогою статистичної інформації, виділеної з побудованих керуючих просторів великих корпусів природньомовних текстів. Коротко описано відомі підходи до вирішення проблеми анафори та запропоновано підхід до синтезу алгоритму Міткова з запропонованим алгоритмом з метою покращення якості їх роботи. Продемонстровано принцип роботи алгоритму на простому прикладі, що вказує на можливості алгоритму вирішувати неоднозначності семантичного характеру. Подальша робота полягає в реалізації алгоритму, тренуванні коефіцієнтів та отриманні показників якості в термінах покриття та точності (recall/precision).

Список використаних джерел

1. *Shalom Lappin, Herbert J. Leass* An Algorithm for Pronominal Anaphora Resolution// Computational Linguistics.– Volume 20.– Issue 4.– December 1994.– pp. 535-561.
2. *J. Hobbs* Resolving Pronoun References// Lingua.– Vol. 44.– pp. 311-338.
3. *Mitkov R.*, Anaphora resolution, London:Longman, 2002, 240 pages.
4. *Анисимов А.В.* Управляющее пространство синтаксических структур естественного языка// Кибернетика.– 1990.– №3.– С. 11-17.
5. *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.
6. *Anisimov A.V., Marchenko O.O., Taranukha V.Yu., Vozniuk T.G.* Development of a semantic and syntactic model of natural language by means of non-negative matrix and tensor factorization// Lecture Notes in Computer Science.– 2014.– preprint.
7. *Kirkpatrick S., Gelatt Jr C. D., Vecchi M. P.,* Optimization by Simulated Annealing// Science .–220 (4598): 1983.– pp. 671–680.

References

1. S. LAPPIN, H. J. LEASS (1994), *An Algorithm for Pronominal Anaphora Resolution*, Computational Linguistics:20(4), pp. 535-561.
2. J. HOBBS (1978), *Resolving Pronoun References*, Lingua, Vol. 44, pp. 311-338
3. MITKOV R. (2002), *Anaphora resolution*, London:Longman, 240 pages.
4. ANYSYMOV A.V (1990), *Upravliaiushchee prostranstvo syntaksycheskykh struktur estestvennoho yazika*, Kibernetyka 3, pp. 11-17.
5. VOZNIUK T.G.(2014), *An algorithm for constructing the control space of syntactic structures of natural language sentences*, Bulletin of Taras Shevchenko National University of Kyiv, Series Physics & Mathematics, #1, pp. 122-127.
6. ANISIMOV A.V., MARCHENKO O.O, TARANUKHA V.Y, VOZNIUK T.G. (2014), *Development of a semantic and syntactic model of natural language by means of non-negative matrix and tensor factorization*, Lecture Notes in Computer Science, preprint.
7. KIRKPATRICK S., GELATT JR C.D., VECCHI M.P. (1983), *Optimization by Simulated Annealing*, Science 220 (4598): pp. 671–680.

Надійшла до редколегії 17.04.14