УДК 681.3.062

Вознюк Т.Г., аспірант.

Побудова класифікатора для вирішення займенникової анафори на основі тензорної моделі

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

T.G. Vozniuk, Postgraduate Student.

Building a classifier to solve pronominal anaphora based on the Tensor model

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

В роботі запропоновано підхід до вирішення займенникової анафори за допомогою побудови класифікатора, який на основі даних про синтаксичні та семанитчні властивості слів анафори та антецедента робить висновок про їх сумісність чи не сумісність.

Ключові слова: метод опорних векторів, комп'ютерна лінгвістика, керуючий простір, проблема анафори.

The paper presents an approach for solving the pronominal anaphora. A short description of existing approaches to resolve the anaphora using classifiers and semantic knowledge was given. There was formulated twenty five heuristic syntactic criteria. Some of them take into the account the morphological properties of the anaphora and the antecedent. Other are based on the distance in words in general and distance in specific part of speech between the two. Criteria based on evaluation of the compatibility by the method of support vectors machine to the already recognized anaphora requires modification to training algorithm, because in this case training process can change training data. Eight criteria were obtained by modifying the estimation algorithm based on the Tensor Model of natural language texts. The basis of the syntactic component analysis with tensor model-based algorithm is control spaces of natural language sentences. Semantic analysis was done using the semantic ontology WordNet and the factorized tensor of typical control spaces. Proposed classifier builds hyperplane in thirty-three dimensional space between the points of matching and not matching pairs of anaphora. Problem of finding this hyperplane has been reduced to an optimization problem. Its dual problem can be solved by gradient descent.

Keywords: support vector machine, computational linguistics, control space, anaphora resolution.

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

В наш час інформаційні технології все більше проникають в повсякденне життя людей. Більшість людей які користуються даними технологіями не є фахівцями в цій галузі, а є пересічними людьми різних професій. Тому інтерфейси програмних додатків стають все більш простими. Один з найбільш зрозумілих інтерфейсів для людини – це засоби спілкування природньою мовою. З одного боку це розробки, що дозволяють синтезувати звук з тексту, тобто читати текст в голос (також розвиваються продукти, що займаються протилежною задачею – розпізнавання звуків природньої мови, тобто «слухати» людей), з іншого боку необхідні засоби аналізу та синтезу природньомовних текстів. Останні дослідження пов'язані з синтаксичним аналізом текстів показують високі результати. Тоді як семантичний аналіз текстів містить в собі багато не вирішених проблем. Одна з них – розв'язання анафори. Можна виділити наступні типи анафори : займенникова, іменникова, прислівна та нульова. Метою даної статті є покращення результатів вирішення займенникової анафори.

Серед робіт присвячених розрізненню займенникової анафори слід виділити алгоритм Міткова[1]. Представлена в його роботі модель узагальнює багато існуючих на той час еврістик в одній зручній формі. А саме виділення деяких критеріїв оцінювання антецедентів по відношенню до анафори на основі синтаксичної та морфологічної інформації, надання кожному

критерію деякої ваги і сумування ваг критеріїв, яким задовільняє дана пара.

В роботі [2] представлений алгоритм, що дозволяє покращити роботу алгоритму Міткова за допомогою використання семантичної інформації.

В роботі [3] описаний критерій для алгоритму Міткова заснований на використанні семантичної інформації представленої тензорною моделлю.

В роботі [4] було побудовано класифікатор, який на основі синтаксичної інформації дає відповідь про сумісність чи несумісність пари антецедент-анафора.

Дана робота являє собою синтез підходів описаних в [3] та [4], тобто буде побудовано класифікатор, що враховує семантичну інформацію.

Критерії оцінювання на основі синтаксичної інформації

Серед критеріїв оцінювання можна виділити ті, що приймають бінарне, натуральне та дійсне значення.

Бінарні критерії кодуються стандартно: 0 – критерій не виконується, 1 – критерій виконується. Наведемо перелік використаних бінарних критеріїв:

1. Антецедент є підметом речення.
2. Антецедент є власною назвою.

3-11. Бінарні ознаки чоловічого, жіного, середнього роду, множини анафори та антецеденту.

Розглянемо критерії, що представлені натуральними числами.

12. Кількість входжень антецедента в текст.
13. Кількість входжень антецедента в дане речення.
14. Кількість використань антецедента в анафоричних зв’язках.
15. Кількість власних назв між анафорою та антецедентом.
16. Кількість іменникоів між анафорою та антецедентом.
17. Кількість займенників між анафорою та антецедентом.
18. Кількість речень між анафорою та антецедентом.
19. Порядковий номер серед іменників антецедента в реченні.
20. Порядковий номер серед займенників анафори в реченні.
21. Кількість анафор, для яких було раніше визначена відповідність даному антецеденту на проміжку між антецедентом та анафорою.
22. Кількість анафор, для яких було раніше визначена не відповідність даному антецеденту на проміжку між антецедентом та анафорою.

Найскладнішими є критерії, значення яких представлене дійсним числом.

23. Кандидат входить в альфа-звязок від слова, що знаходиться в одній і тій же фактор-множині як і слова discuss, illustrate, identify тощо. Для перевірки критерію необхідно взяти значення з матриці альфа-бета зв’язків по парі слів (identify, антецедент). Також треба зробити запит до таблиці трійок за трійкою (антецедент, identify, -). Вибір іншого слова (illustrate чи discuss) не суттєво вплине на результати, оскільки всі вони будуть знаходитися в одній фактор множині.
24. Середнє значення оцінки пари антецедент-анафора методом опорних векторів, для яких було раніше визначена відповідність на проміжку між антецедентом та анафорою.
25. Середнє значення оцінки пари антецедент-анафора методом опорних векторів, для яких було раніше визначена не відповідність на проміжку між антецедентом та анафорою.

Критерії 21, 22, 24 та 25 можуть бути введені в процес тренування тільки після початкового тренування на всіх інших критеріях, оскільки вони вимагають вже існуючого класифікатора. Крім того після кожного перетренування з врахуванням цих додаткових критеріїв вид класифікатора змінюється. Тому перерахунок результатів цих критеріїв необхідно включити в процес тренування.

Модифікація тензорного алгоритму оцінювання пари антецедент-анафора

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

Сформулюємо алгоритм підрахунку семантико-синтаксичних критеріїв для заданої пари w та w'.

I. Побудувати керуючі простори cs та cs' для речень s та s' за допомогою алгоритму запропонованого в [5].

II. 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

III. 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'

IV. Обчислити бінарний критерій:

$$\begin{cases} 1, & \text{якщо two.type= two'.type} \\ 0, & \text{інакше} \end{cases}$$

V. Обрахувати натуральнозначні критерії на основі отриманої інформації та знаходження семантичної відстані між словами по онтології WordNet[6].
   1. two.first та two'.first
   2. two.second та two'.second
   3. three.first та three'.first
   4. three.second та three'.second
   5. three.third та three'.third

VI. Знайти дійснозначні критерії вживаності антецедента в контексті анафори за допомогою запитів до тензорної моделі за наступними векторами:
   1. Якщо two'.first = w',
      то (w, two'.second)
      інакше (two'.first, w)
   2. Якщо three'.first=w'
      то (w, three'.second, three'.third)
      інакше якщо three'.two=w'
      то (three.first, w, three'.third)
      інакше (three'.first, three'.second, w)

Критерій IV є аналогом критерію синтаксичного паралелізму, оскільки він перевіряє, чи в однаковий тип зв'язку входить антецедент та анафора.

Критерії V.1-5 відповідають за синтаксично-семантичний паралелізм та перевіряють чи схожі за змістом слова знаходяться в однакових синтаксичних відношеннях відносно анафори та антецедента.

За допомогою критеріїв VI.1-2 за допомогою тензорної моделі мови визначається, чи можна вжити антецедент в реченні замість анафори. Якщо при заміні анафори антецедента в парі та трійці утворюються семантично не корректні фрази, то така пара повинна бути відкинута.

Таким чином ми сформували 8 критеріїв на основі аналізу контексту антецедента та анафори за допомогою тензорної моделі мови

Тренування класифікатора методу опорних векторів

Скористаємося навчанням з вчителем (supervised learning). Для даного типу навчання нам необхідний розмічена тренувальна вибірка. На основі тренувальної вибірки треба сформувати позитивні та негативні приклади для навчання. Для цього обрахуємо сформульовані вище ознаки всіх пар антецедентів та анафор. Кількість пар в негативних прикладах буде близькою до n*m, де n – кількість анафоричних займенників в тренувальній вибірці, m – кількість кандидатів в антецеденти. Тому кількість негативних ознак буде значно більше ніж позитивних. Для пришвидшення навчання можна зменшити кількість негативних прикладів, викинувши з тренуючої вибірки пари негативні пари антецедент-анафора, відстань між якими більша ніж деяке порогове значення, вибране виходячи з ступеня зв'язності тренуючої вибірки. Маємо:

$$D=\{(x_i,y_i)\,|\,x_i\in\Re^{33},y_i\in\{-1,1\}\}_{i=1}^{n}$$

де x_i - і-та точка тренувальної вибірки в просторі ознак,

$$y_i=\begin{cases} 1, & \text{якщо тточк представляє сумісну пару} \\ -1, & \text{інакше} \end{cases}$$

,

n – кількість пар в тренувальній вибірці.

В лінійному випадку методу опорних векторів поверхня має вигляд гіпер-площини. Цю гіперплощину можна задати наступним рівнянням:

$$w\cdot x-b=0,$$

де $\cdot$ – скалярний добуток.

Сучасні алгоритми штучного інтелекту не дійшли до повного розуміння тексту. Сформульований вище простір ознак не є виключенням. Тому деякі точки в просторі можуть вказувати на сумісну або не сумісну в деякому контексті пару. Це означає, що побудована для класифікатора поверхня не зможе розділити простір на 2 частини, в кожній з яких будуть точки свого класу. Додатковою проблемою нероздільності класів може бути не правильно обраний вид поверхні, що їх розділює. В нашому випадку – це гіперплощина. Для нероздільних класів використовується модифікація алгоритму опорних векторів з м'яким зазором. В цьому випадку тренування полягає в вирішенні наступної оптимізаційної задачі:

$$\arg\min_{w,b,\xi} \left(\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i\right)$$

$$y_i \cdot (w \cdot x_i - b) \geq 1 - \xi_i, 1 \leq i \leq n$$

$$\xi_i \geq 0$$

де С – лінійний штрафний коефіцієнт

Така задача може бути вирішена будь-яким методом оптимазації.

Висновки

У роботі було представлено метод вирішення проблеми анафори за допомогою тренування класифікатора. Було сформовано 33-вимірний простір ознак. Серед них, 8 ознак використовують семантико-синтаксичну інформацію представленою тензорною моделлю мови.

Список використаних джерел

1. Mitkov R. A New, Fully Automatic Version of Mitkov's Knowledge-Poor Pronoun Resolution Method // CICLing. — 2002. — PP. 168—186.
2. Марченко О.О. Semantic Modification of the Mitkov Algorithm for Anaphora Resolution // Штучний інтелект. — 2012 — № 3. — С. 106—110.
3. 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. 100—103.
4. Алгоритм автоматизированного разрешенеия анафоры местоимений третьего лица на основе методов машинного обучения [Электронный реурс] / Толпегин П.В., Ветров Д.П, Кропотов Д.А. // Диалог — 2006 — Режим доступа : http://www.dialog-21.ru/digests/dialog2006/materials/html/Tolpegin.htm (20.05.14). — Загл. с экрана
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. Марченко О.О. Методи оцінювання семантичної близькості–зв'язності слів природної мови // Штучний інтелект. — 2012 — №4 — С. 213—219.

References

1. MITKOV, R. (2002) Fully Automatic Version of Mitkov's Knowledge-Poor Pronoun Resolution Method, In *CICLing*, 2002. pp. 168-186.
2. MARCHENKO, O.O (2002) Semantic Modification of the Mitkov Algorithm for Anaphora Resolution, *Artificial intelligence* #3, pp. 106-110.
3. VOZNIUK, T.G.(2014) Application of the control space of syntactic structures for anaphora resolution, *Bulletin of Taras Shevchenko National University of Kyiv Series Physics & Mathematics*, #2, pp. 100-103
4. TOLPEHIN, P.V & VETROV, D.P & KROPOTOV D.A.(2006) Algorithm for automatic pronominal anaphora resolution based on machine learning, *Dialog*, [Online] Available from http://www.dialog-21.ru/digests/dialog2006/materials/html/Tolpegin.htm.
5. VOZNIUK, T.G.(2014) Algorithm for construction of the control space of syntactic structures, *Bulletin of Taras Shevchenko National University of Kyiv Series Physics & Mathematics*, #1, pp. 122-127.
6. MARCHENKO O.O.*(2012*) Methods for estimation of the semantic distance of natural language words, *Artificial intelligence*, №4, pp. 213- 219.

Надійшла до редколегії 18.06.14