Список літератури

1. Dietterich T. G. Solving Multiclass Learning Problems via Error-Correcting Output Codes / T. G. Dietterich, G. Bakiri // Artificial Intelligence Research. – Vol. 2. – 1995. – P. 263–286.
2. Pujol O. Discriminant ECOC: A Heuristic Method for Application Dependent Design of Error Correcting Output Codes / O. Pujol, P. Radeva, J.Vitriá // IEEE Transaction on pattern analysis and machine intelligence. – Vol. 28. – № 6. – 2006. – P. 1107–1012.
3. UCI Machine Learning Repository. Dermatology Data Set [Електронний ресурс]. – Режим доступу: http://archive.ics.uci.edu/ml/datasets/Dermatology. – Назва з екрана.

O. Porkhun

DETERMINING THE DERMATOLOGY DISEASES USING THE DISTRIBUTED OUTPUT CODE MODELS AND PERCEPTRON

In the article the distributed output code models for solving multiclass learning problems with usage multilayer perceptron are considered and application of the developed multi-classification system implementing the described models for determining the dermatology diseases is described.

Keywords: multi-classification, distributed output code, exhaustive codes, matrix of codewords, multilayer perceptron.

Матеріал надійшов 10.06.2013

УДК 681.3

Марченко О. О.

ПОБУДОВА ЛЕКСИКО-СИНТАКСИЧНОЇ МОДЕЛІ ПРИРОДНОЇ МОВИ ІЗ ЗАСТОСУВАННЯМ СУЧАСНИХ МЕТОДІВ ОБРОБКИ ВЕЛИКИХ ТЕКСТОВИХ КОРПУСІВ

Статтю присвячено розробці алгоритму формування моделі лексико-синтаксичних структурних зв'язків природної мови на основі частотно-синтаксичного аналізу речень великого текстового корпусу. Для запису універсальних структур необмеженої складності та довжини використано модель керуючих просторів синтаксичних структур речень природної мови. Для ефективного та економного представлення даних розріджені масиви трансформовано за допомогою методів невід'ємної факторизації матриць та тензорів.

Ключові слова: обробка текстів природною мовою, керуючі простори синтаксичних структур, невід'ємна факторизація тензорів.

Вступ

Разом із значним зростанням обчислювальної потужності сучасних комп'ютерів і появою нових інтелектуальних алгоритмів обробки великих масивів інформації останнім часом розробка нових методів розв'язку багатьох задач штучного інтелекту вийшла на якісно новий рівень. Серед таких фундаментальних алгоритмів обробки великих масивів інформації виділяється універсальний і потужний підхід – невід'ємна тензорна факторизація.

Невід'ємна тензорна факторизація сьогодні широко затребувана в таких областях, як машинне навчання, обробка зображень, інформаційний пошук, обробка природної мови, та інших напрямках. Такий підхід є одним з найбільш перспективних для виявлення та аналізу взаємозв'язків і від-

ношень у даних, які містять сполучення об'єктів N різних типів і класів. N-мірний тензор, що в інформатиці трактується як багатовимірний масив даних, є зручною структурою для представлення даних вищих порядків. Факторизація N-мірного тензора при розкладанні генерує N матриць, які складаються з векторів-стовпчиків розмірності k. Ці N матриць представляють відображення кожного виміру тензора на k фактор-вимірів прихованого семантичного простору. Це служить унікальним засобом для моделювання та виявлення взаємозв'язків і спільної поведінки N змінних у масиві N-мірних даних. Факторизація тензора є мультилінійним аналогом сингулярного розкладання матриць, який використовується в латентному семантичному аналізі для обробки двовимірних масивів даних. У певному сенсі можна назвати метод невід'ємної факторизації тензорів n-мірним узагальненням латентного семантичного аналізу. Структуру, отриману в результаті факторизації тензора, можна порівняти з багатошаровою нейронною мережею, що складається з N шарів, які представляють множини об'єктів N типів, і з прихованого комутаційного шару, який містить множину вузлів з різними ваговими коефіцієнтами і моделює взаємозв'язки між об'єктами N типів, тобто пов'язує дані N шарів у єдину нейронну мережу.

Невід'ємна тензорна факторизація є перспективним методом у вирішенні задач комп'ютерної лінгвістики, про що свідчать численні роботи в цьому напрямку [8; 9; 10; 11].

Особливий інтерес представляють роботи [9; 10], в яких описано моделі тензорного представлення даних про частоту різних типів синтаксичних сполучень слів у реченнях, наприклад 3-мірних сполучень типу subject – verb – object, 4-мірних сполучень типу subject – verb – direct_object – indirect_object, або інших синтаксичних сполучень довжини, що не перевищує розмірність тензора N. У тензорі кожен вимір відповідає деякому фіксованому члену речення – підмет, присудок, додаток, означення, обставина та інші. N-мірні тензори містять оцінки частоти вживання сполучень різних наборів слів у реченнях природної мови, де враховуються синтаксичні позиції слів у реченнях. Після обробки великих текстових корпусів та накопичення значного обсягу даних у тензорі, формується N-вимірний масив опису поведінки лексичних одиниць у реченнях даної мови: тобто для множини слів у тензорі дано опис, у які синтаксичні відношення ці слова мають властивість вступати, з якими словами вони вступають у ці відношення і з якою частотою. Причому, відношення ці не бінарні, а багатомірні (N - максимальна розмірність відношень). Наступним є етап невід'ємної факторизації отриманого тензора. Факторизація призводить до значного перетворення моделі представлення даних. Спочатку багатовимірний тензор є розрядженим і надвеликим за обсягом. Кожен з N вимірів синтаксичного простору містить десятки тисяч чи сотні тисяч точок – слів. Після факторизації тензора його дані представлено у вигляді N матриць, що складаються з векторів-стовпчиків розмірності k (де значення k набагато менше, ніж кількість точок-слів у будь-якому з N вимірів тензора). Параметр k – це розмірність латентного семантичного простору, число ознакових вимірів у ньому. Після цього, крім значно компактнішого та зручнішого представлення масиву даних, з'являється можливість швидкого обчислення оцінки ймовірності будь-якого можливого сполучення слів у різних синтаксичних конструкціях речення. Це можна зробити шляхом обчислення суми добутків компонент N k-мірних векторів-стовпчиків, що відповідають цим словам, з матриць, що відповідають їх синтаксичним позиціям. Наприклад, щоб перевірити, наскільки імовірним є речення «Кухар смажить качку», потрібно знайти в матриці SUBJECT k-мірний вектор-стовпчик, який відповідає іменнику «кухар», потім знайти в матриці VERB k-мірний вектор-стовпчик, що відповідає дієслову «смажить», потім знайти в матриці DIRECT_OBJECT k-мірний вектор-стовпчик, який відповідає іменнику «качка»; після цього обчислюється сума добутків відповідних компонент цих трьох векторів, тобто обчислюється $x_{svdo} = \sum_{i=1}^{k} s_{si} v_{vi} do_{doi}$ (для випадку N = 3).

Якщо результат суми перевищує деякий пороговий рівень, то робиться висновок про можливість існування такої послідовності слів у реченні. Обчислення даної оцінки для словосполучення «Качка смажить кухаря» призведе до висновку про неможливість такого варіанта.

Дана модель дозволяє досить успішно автоматично виділяти з корпусів текстів такі лінгвістичні структури, як ***селективні преференції*** в реченнях (selectional preferences) [9] та ***субкатегоріальні фрейми дієслів*** (Verb SubCategorization Frame) [10], які поєднують у собі дані про семантичні та синтаксичні властивості зв'язків-відношень між дієсловами та їх аргументами-іменниками в реченнях природної мови.

Недоліком такої моделі можна вважати негнучкість і обмеженість представлення синтаксису. Розмірність тензора обмежує максимальну

довжину речень-словосполучень, що описуються даною моделлю. Кожній осі відповідає конкретна синтаксична позиція. У роботі [9] описано 3-вимірний тензор для моделювання одного синтаксичного сполучення – підмет-присудок-додаток. У роботі [10] автор описує тензори розмірності 9 і 12 для моделювання двох десятків різних типів синтаксичних відношень-сполучень. Просте збільшення розмірності тензора для обробки більшої кількості типів синтаксичних відношень розширеної арності не виглядає надто переконливим засобом вдосконалення моделі. Актуальним і затребуваним напрямом досліджень є аналіз і розробка універсальних засобів завдання синтаксичних структур речень природної мови. Необхідно використовувати формальну модель представлення, яка за допомогою рекурсії могла б виразити синтаксичні відношення речень довільної довжини і будь-якого ступеня складності структури і дозволила б записати структурний зв'язок між словами в реченнях будь-якої довжини у тензорі фіксованої розмірності. Як подібну модель представлення синтаксису мови пропоновано використати **керуючий простір синтаксичних структур природної мови** [1].

1. Керуючий простір синтаксичних структур природної мови

Основні синтаксичні конструкції описано в класичних схемах граматики мови, які відносяться до періоду античності і мало змінилися до теперішнього часу.

Досить тонкі відношення керування між словами виражаються у відомих лінгвістичних моделях дерев підпорядкування і систем складових, що з'явилися в 1950 роках. Очевидною перевагою перерахованих моделей є їх коректність – адекватне відображення тих чи інших специфічних характеристик синтаксичної структури речення. Причому дані моделі не позбавлено недоліків. Модель дерев підпорядкування орієнтована на керуючі зв'язки тільки між словами, а модель систем складових враховує ієрархічне відношення вкладеності словосполучень в лінійній структурі тексту. Ці чинники лише наближено описують дійсні комунікативні властивості, що містяться в синтаксичних структурах.

Спроби побудови зручніших для машинної обробки моделей, узагальнюючих властивості дерев підпорядкування і систем складових призвели до створення моделі системи компонент А. С. Нариньяні [3] та синтаксичних груп А. В. Гладкого [2]. У цих моделях відбувається переміщення кута зору на синтаксичні структури з лінійного порядку, нав'язаного послідовністю запису тексту, до складного простору, утвореного синтаксично зв'язаними групами об'єктів. А. В. Анісімов запропонував перейти до простору представлення, не залежного від порядку запису тексту, отже, і від національної мови, який виражає всі предикативні та синтагматичні відношення, що містяться в синтаксичних структурах. Такий простір назвали керуючим [1].

Розглянемо запропоновану алгоритмічну модель речення природної мови. На відміну від суто лінгвістичного підходу, речення розглядається як певний динамічний обчислювальний рекурсивний процес, що розвивається в керуючому просторі, який пов'язує синтаксично згруповані частини речення інформаційними каналами. Структура керуючого простору відображає семантику синтагматичних і предикативних конструкцій мови.

Крім властивості давати імена об'єктам навколишнього світу, мова володіє фундаментальною властивістю виражати динамічні відношення, в які вступають об'єкти. Так, дієслово пов'язує у відношення об'єкти, що беруть участь у схемі дії цього дієслова, прикметник задає відношення об'єкта з самим собою. З синтаксичної моделі ми повинні знати, які частини речення пов'язані між собою через відношення і якого типу ці відношення. Існують два види синтаксичних відношень – предикативні і синтагматичні. Предикативне відношення виражає залежність між синтаксичними об'єктами через поняття, що означає дію і, зазвичай, виражається за допомогою присудка. Синтагма – це поєднання двох синтаксичних об'єктів, з яких один є визначенням іншого, тому в моделі повинні повністю виражатися саме ці види відношень. Крім того, в широкому розумінні синтагми повинні утворювати синтаксичні групи.

Адекватна модель синтаксичної структури повинна також відображати основну властивість рекурсивності мови – здатність розгортати власні визначення, тобто давати уточнення, характеристики, коментарі до своїх частин, а також будувати визначення визначень.

Перш за все треба зупинитися на ролі керуючих зв'язків між словосполученнями. Навмисно порушується традиційний лінгвістичний підхід, при якому присудок вважається головним членом речення, від якого ідуть керуючі зв'язки. Це успадкувалося від звички вважати ім'я функції головнішим, аніж її аргументи. Для побудови даної моделі зручніше задавати синтаксичні відношення зв'язками *генерації* і *передачі відношень*. При цьому досягається точніша характеристика

керуючих зв'язків у порівнянні з традиційним підходом.

Якщо два об'єкти A і B вступають у відношення C, то ми виділяємо об'єкт (припустимо, A), що викликає (ініціює, породжує) це відношення C, і об'єкт, на який передається це відношення, – B. Таким чином, виділяємо два види спрямованих зв'язків: від *об'єкта-генератора відношення* до *відношення* і від *відношення* до підпорядкованого об'єкта. Перший вид зв'язку називаємо α-зв'язком (зв'язок генерування), другий – β-зв'язком (зв'язок розповсюдження). Об'єкти A, B і відношення C розміщено в точках керуючого простору і тому графічне представлення відношення C, що зв'язує A і B, має вигляд, зображений на рисунку 1.

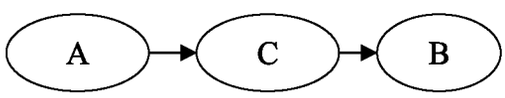

Рис. 1. Об'єкт A генерує відношення C, яке передається на об'єкт B

Дієслова визначають відношення між об'єктами. Тому в стандартній схемі простого речення: «іменник – дієслово – іменник» α-зв'язок спрямований від першого іменника до дієслова, а β-зв'язок спрямований від дієслова до іменника-визначення. Розглянемо приклад: *Хлопчик купив м'яч*. Об'єкт *хлопчик* генерує відношення *купив* і направляє його на об'єкт *м'яч*. Тому α-β-структура цього речення має вигляд, як показано на рисунку 2.

Рис. 2. Структура речення *Хлопчик купив м'яч*

Розглянемо фразу: *Високий хлопчик*. Тут об'єкт *хлопчик* генерує унарне відношення *високий* і передає це відношення собі, як показано на рисунку 3. Виникає кільцевий зв'язок, що характеризує визначення.

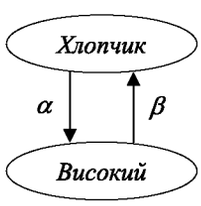

Рис. 3. Структура словосполучення *Високий хлопчик*

Аналогічно міркуючи, для фрази *Високий хлопчик дуже любить футбол* отримуємо таку структуру, як показано на рисунку 4.

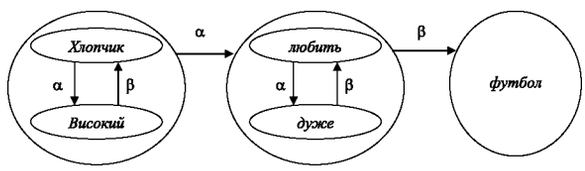

Рис. 4. Структура речення *Високий хлопчик дуже любить футбол*

Речення мають два типи α-β-зв'язків: строго лінійна залежність і замкнута кільцева залежність. Першу називають лінійною конструкцією, другу – визначенням. Перша відповідає предикативним конструкціям мови, друга – синтагматичним.

Формальна модель, орієнтована на завдання складних структур необхідного виду у формі керуючих просторів, будується таким чином.

Дано клас базових об'єктів U, що містить порожню множину. З кожним об'єктом асоціюється певний тип. Всього різних типів скінчене число. Типи можна виразити числами з інтервалу [0, N]. Припускаємо неоднозначність при зіставленні об'єктів типам, тобто функція приписування типів φ відображає U у множину всіх підмножин, утворених числами з інтервалу [0, N]. Конструкціями є або об'єкти з U, або конструкції, що отримані з інших конструкцій за допомогою підстановки останніх в точки лінійної або визначальної залежності. Правила обчислення типів конструкцій мають такий вигляд.

1. Якщо в лінійній залежності об'єкт типу i з'єднується α-зв'язком з об'єктом типу j, а останній β-зв'язком з об'єктом типу k, то тип такої конструкції дорівнює f (i, j, k), де f – задана частково-визначена функція.

2. Нехай d (i, j) – задана частково-визначена функція, тотожно рівна 1 в точках свого визначення. Функція d називається функцією узгодження.

Якщо об'єкт типу i у визначальній конструкції уточнюється за допомогою об'єкта типу j, то всій конструкції приписується тип i, якщо тільки значення функції узгодження дорівнює 1. В іншому випадку значення типу не визначено.

Оскільки множина базових типів є скінченною, то функції f і d можуть бути задані таблицями. Правило 2 дозволяє легко обчислювати тип будь-якої складної конструкції. Крім того, легко довести, що тип будь-якої конструкції збігається з типом однієї з базових конструкцій, що задаються функціями f або d.

Якщо неможливо обчислити тип конструкції, то вона вважається некоректною. Всі коректні конструкції утворюють керуючі простори класу U.

Щодо синтаксичних структур дане визначення уточнюється таким чином.

Базові об'єкти – це слова і прості словосполучення, що представляють частини мови (іменники, прикметники, дієслова, частки і т.д.) з відповідними морфологічними ознаками, а також складносурядні відношення і корелятори, призначені для з'єднання підпорядкованих речень з головними. Тип слова – це повна його граматична характеристика. Наприклад, тип слова *книга* дорівнює (іменник, неживе, однина, називний відмінок). Можливе розширення поняття типу додаванням деяких семантичних атрибутів. Неоднозначність завдання типу проявляється в неоднозначності розуміння значення деяких слів у відриві від контексту. Наприклад, слово *мати* може бути іменником або дієсловом. Функція f задає типи простих речень, а також тип складного речення залежно від його конструкції верхнього рівня. До базових об'єктів відносимо корелятори, що являють або скріпу, або пару службових слів – (скріпа, співвідносне слово). Скріпа – службове слово в підпорядкованому реченні, що служить для прив'язування цього речення до головного. Співвідносне слово знаходиться у головному реченні (якщо воно є) і служить для зв'язку з відповідною скріпою. Детальну класифікацію кореляторів виконано в [2]. Функція d задає умови узгодження типів об'єкта, що визначається, та об'єкта, який визначає. Наприклад, визначеннями до іменника можуть бути прикметники, прийменникові групи або підпорядковані речення, до дієслова – прислівник, дієприслівник або підпорядковане речення; дієслово не може бути визначенням для іменника і таке інше.

Отже, функції f і d виконують роль фільтра, що виділяє допустимі конструкції. Побудова таблиць значень функцій f і d не завдасть великих труднощів. Усе необхідне для цього є в класичній граматиці мови. Оскільки у визначальній конструкції роль підпорядкованої частини зводиться до коментарю або уточнення головної частини, то значення типу всієї синтагматичної конструкції вибрано рівним значенню головного об'єкта – генератора властивості.

При обробці синтаксичних структур природної мови виникають два основні завдання – аналіз і синтез. Завдання аналізу – за наявним лінійним записом тексту побудувати відповідний КП, завдання синтезу – за наявним КП побудувати відповідну лінійну послідовність тексту. Друге завдання простіше. При відображенні лінійної залежності в текст алгоритм синтезу слідує послідовності α-β-зв'язків. При синтезі тексту з визначень спочатку вирішується, що видається раніше – головний об'єкт або його уточнення. Після вирішення цього питання відбувається видача тексту з відповідних конструкцій, які утворюють визначення. В алгоритмі аналізу йде багаторазова згортка визначень і виділення лінійних залежностей.

2. Побудова лексико-синтаксичної моделі природної мови

Для побудови семантико-синтаксичної моделі природної мови пропоновано розробити систему автоматичного заповнення тривимірного тензора функції F та матриці функції D в процесі синтаксичного аналізу та пост-обробки речень великого текстового корпусу. Система повинна виконувати таку послідовність перетворень лінгвістичних структур:

1) система послідовно приймає на вхід речення з великого тактовного корпусу та виконує їх синтаксичний аналіз за допомогою модуля граматичного розбору Stanford Parser [12], який генерує синтаксичну структуру речень у вигляді дерев підпорядкування;

2) програма аналізує дерево підпорядкування поточного речення, збираючи керуючий простір його синтаксичної структури, перебираючи зв'язки між словами для виявлення предикативних сполучень довжиною 3 (типу *підмет-присудок-додаток* та ін.), а також синтагматичних сполучень довжиною 2 (*іменник-прикметник, дієслово-прислівник* та ін.);

3) після збірки керуючого простору синтаксичної структури поточного речення для кожної трійки точок (i, j, k) у просторі, зв'язаних лінійною послідовністю α-β-зв'язків, до тензора F у комірку F [I, J, K] заноситься значення f (i, j, k) – $F[I, J, K] := f(i, j, k)$. Координати I, J, K відповідають лексичному значенню точок (за кожною точкою керуючого простору фіксується певне слово чи словосполучення), також містять у закодованому вигляді інформацію про граматичну категорію даних точок (частина мови та ін.);

4) аналогічно для кожної пари точок (i, j) у керуючому просторі синтаксичної структури поточного речення, зв'язаних між собою кільцевим синтагматичним α-β-зв'язком, до матриці D у комірку D [I, J] заноситься значення d (i, j) – $D[I, J] := d(i, j)$. Координати I, J відповідають лексичному значенню точок, а також містять у закодованому вигляді інформацію про граматичну категорію даних точок (частина мови та ін.).

Після накопичення великої кількості даних у матриці D та у тривимірному тензорі F формується достатній опис табличних функцій F та D для ефективної реалізації лексико-синтаксичної моделі природної мови на основі керуючого простору синтаксичних структур природної мови. Надвелика розмірність та розрідженість утвореної матриці D та побудованого тензора F вимагають трансформації структур даних з метою економнішого та зручнішого представлення для збереження, обробки та редагування даних. Для оптимізації величезних масивів даних найкраще підходять методи невід'ємної матричної та тензорної факторизації.

3. Факторизація матриці D

Для розкладення матриці великої розмірності D (N × M) у вигляді добутку двох матриць W (N × k) × H (k × M), де (k<<N, M), доцільно використати алгоритм невід'ємної матричної факторизації NMF, що був запропонований Лі та Суном [6]. У якості цільової функції використовується норма Фробеніуса, яка описується формулою 1:

$$\min_{W,H} \|D - WH\|_F^2 \quad (1),$$

причому елементи матриць W та H повинні бути невід'ємними.

Для такої цільової функції та для двох початкових матриць W_0 і H_0, NMF алгоритм складається з ітераційного виконання двох кроків:

$$1.\ (H_k)_{i,j} = (H_{k-1})_{i,j} \times \frac{(W_{k-1}^T D)_{i,j}}{(W_{k-1}^T W_{k-1} H_{k-1})_{i,j}} \quad (2).$$

$$2.\ (W_k)_{i,j} = (W_{k-1})_{i,j} \times \frac{(D H_{k-1}^T)_{i,j}}{(W_{k-1} H_{k-1} H_{k-1}^T)_{i,j}} \quad (3).$$

На практиці кроки алгоритму повторюються, доки не буде досягнуто нерухомої точки або не буде виконано максимальну кількість ітерацій. Лі та Сун довели дві основні властивості цього алгоритму: по-перше, цільова функція є монотонно спадною під час застосування правил; по-друге, матриці W і H стають постійними тільки у випадку досягнення стаціонарної точки цільової функції.

4. Факторизація тензора F

Для розкладання тензора використовується метод невід'ємної тензорної факторизації. Він подібний до паралельного факторного аналізу з обмеженням, що всі дані мають бути невід'ємними. Паралельний факторний аналіз – це мультилінійний аналог сингулярного розкладання матриць, що використовується в латентному семантичному аналізі [7]. Головна ідея методу – мінімізація суми квадратів різниць між оригінальним тензором і факторизованою моделлю тензора. Для 3-вимірного тензора $T \in R^{D_1 \times D_2 \times D_3}$ визначається цільова функція (4), де k – розмірність факторизованої моделі, а ∘ – зовнішній добуток (outer product) :

$$\min_{x_i \in R^{D_1}, y_i \in R^{D_2}, z_i \in R^{D_3}} \| T - \sum_{i=1}^{k} x_i \circ y_i \circ z_i \|_F^2 \quad (4).$$

Для невід'ємної факторизації додаються обмеження по невід'ємності значень елементів (5) :

$$\min_{x_i \in R_{\geq 0}^{D_1}, y_i \in R_{\geq 0}^{D_2}, \ldots, z_i \in R_{\geq 0}^{D_N}} \| T - \sum_{i=1}^{k} x_i \circ y_i \circ \ldots \circ z_i \|_F^2 \quad (5).$$

Результат роботи алгоритму – представлення тензора у вигляді трьох матриць, які описують відображення кожної з розмірностей тензора на k фактор-вимірів латентного семантичного простору. NTF модель підганяється методом найменших квадратів. На кожній ітерації дві з розмірностей фіксуються, а третя розмірність підганяється методом найменших квадратів. Процес триває до моменту збіжності. Число фактор-вимірів латентного семантичного простору в проведених експериментах було взято k = 150. Виходячи з досвіду попередніх досліджень, саме це значення забезпечує оптимальні результати факторизації [10].

Для розв'язання даної задачі написано програмну реалізацію алгоритму паралельної факторизації PARAFAC [4] тривимірного тензора.

5. Властивості лексико-синтаксичної моделі природної мови

Факторизацією матриці D та тензора F система формує потужну базу знань, яка містить дані про будову синтаксичних структур речень природної мови, в які інтегровано опис лексико-семантичних відношень між словами. Окрім загального синтаксису, що задає структуру речень в загальному абстрактному вигляді, база містить лексико-семантичні обмеження, що визначають, які слова можуть утворювати зв'язок певного синтаксичного типу. Для того щоб визначити, чи можуть два слова ***a*** та ***b*** утворити кільцевий синтагматичний зв'язок, треба взяти вектор-строку W_a матриці W, що відповідає слову ***a***, вектор-стовпчик H_b матриці H, що відповідає

слову ***b***, та обчислити скалярний добуток векторів ($\boldsymbol{W}_a$, $\boldsymbol{H}_b^{\mathrm{T}}$). Якщо значення добутку більше нуля, то даний зв'язок є визначеним. Для того щоб визначити, чи можуть три слова a, b та c утворювати предикативний зв'язок (a → b → c), потрібно з першої матриці X розкладеного тензора F узяти вектор-стовпчик X_a, що відповідає слову ***a***, з другої матриці Y розкладеного тензора F узяти вектор-стовпчик Y_b, що відповідає слову ***b***, з третьої матриці Z розкладеного тензора F узяти вектор-стовпчик Z_c, що відповідає слову ***c***, та обчислити значення $S_{abc} = \sum_{i=1}^{k} X_a[i] * Y_b[i] * Z_c[i]$.

Якщо значення S_{abc} більше нуля, то даний зв'язок є визначеним. Усі зв'язки, які не є визначеними, вважаються невизначеними. Отримані матриці в неявному вигляді задають множину коректних визначених речень мови, що задається текстами вхідного корпусу. Вектори-стовпчики слів з отриманих матриць є неявним описом їх «структурної поведінки» – вони задають, в які синтаксичні відношення ці слова мають властивість вступати та з якими словами вони вступають у ці відношення. За допомогою отриманих матриць можливо виконувати синтаксичний аналіз речень з побудовою керуючого простору їх синтаксичних структур, використовуючи висхідні алгоритми аналізу типу Кока-Янгера-Касамі [5]. Обходом матриць по ненульових елементах вектор-стовпчиків можливо розробити програму синтезу множини речень, визначених побудованою моделлю.

Висновки

Рекурсивність керуючих просторів синтаксичних структур природної мови дозволяє виразити структуру речень довільної складності, довжини та глибини. Це дає можливість при розробці семантико-синтаксичної моделі замість нарощування мірності лінгвістичних масивів сполучності лексичних одиниць обмежитися лише побудовою одного тривимірного тензора та однієї матриці. Із потужності та універсальності описаної моделі випливає її актуальність як в теоретичному плані, так і в аспекті застосування на практиці в прикладних лінгвістичних системах.

Список літератури

1. Анисимов А. В. Управляющее пространство синтаксических структур естественного языка / А. В. Анисимов // Кибернетика. – 1990. – № 3. – С. 11–17.
2. Гладкий А. В. Синтаксические структуры естественного языка в автоматизированных системах общения / А. В. Гладкий // Проблемы искусственного интеллекта. – М. : Наука, 1985. – 144 с.
3. Нариньяни А. С. Формальная модель: общая схема и выбор адекватных средств / А. С. Нариньяни. – Новосибирск : ВЦ СО АН СССР, 1978. – 19 с. – (Препринт № 400/ ВЦ СО АН СССР).
4. Harshman R. Foundations of the PARAFAC procedure: models and conditions for an explanatory multimodal factor analysis / R. Harshman // UCLA Working Papers in Phonetics. – Vol. 16. – 1970. – P. 1–84.
5. Kasami T. An efficient recognition and syntax-analysis algorithm for context-free languages. Scientific report AFCRL-1965–758. – Air Force Cambridge Research Lab. – Bedford. – MA.
6. Lee D. D. Algorithms for Non-Negative Matrix Factorization [Електронний ресурс] / D.D. Lee, H. S. Seung // NIPS. – 2000. – Режим доступу: http://hebb.mit.edu/ people /seung/ papers/nmfconverge.pdf. – Назва з екрана.
7. Indexing by Latent Semantic Analysis / Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, R. Harshman // Journal of the American Society for Information Science. – 1990. – 41 (6). – P. 391–407.
8. Shay B. Cohen Tensor Decomposition for Fast Parsing with Latent-Variable PCFGs / Shay B. Cohen, Michael Collins // NIPS-2012. – P. 2528–2536.
9. Tim Van de Cruys. A Non-negative Tensor Factorization Model for Selectional Preference Induction / Tim Van de Cruys // Journal of Natural Language Engineering. – 2010. – Vol. 16 (4). – P. 417–437.
10. Multi-way Tensor Factorization for Unsupervised Lexical Acquisition / Tim Van de Cruys, Laura Rimell, Thierry Poibeau, Anna Korhonen // Proceedings of COLING 2012. – Mumbai, India. – P. 2703–2720.
11. Wei Peng. On the equivalence between nonnegative tensor factorization and tensorial probabilistic latent semantic analysis / Wei Peng, Tao Li // Applied Intelligence, Springer Journals. – October 2011. – Volume 35, Issue 2. – P. 285–295.
12. The Stanford Natural Language Processing Group [Електронний ресурс]. – Режим доступу: http://nlp.stanford.edu/ software/lex-parser.shtml. – Назва з екрана.

O. Marchenko

DEVELOPMENT OF LEXICAL-SYNTACTIC MODEL OF NATURAL LANGUAGE BY USING MODERN METHODS OF LARGE TEXT CORPORA PROCESSING

Paper considers the algorithm for building the model of lexical and syntactic structural relations of natural language based on frequency-syntactic analysis of sentences from large text corpora. During the analysis the obtained data are accumulated in large arrays. To record syntactic structures of unlimited complexity, depth and length the natural language syntactic structures control spaces were used. The obtained arrays are huge and sparse. For effective and compact storage of the data the arrays were transformed by using the methods of non-negative matrix and tensor factorization.

Keywords: natural language text processing, syntactic structures control spaces, non-negative tensor factorization.

Матеріал надійшов 31.05.2013

УДК 510.6

Касьянюк В. С., Малютенко Л. М.

КЛАСТЕРИЗАЦІЯ ДАНИХ З ВИКОРИСТАННЯМ ТЕОРІЇ МОЖЛИВОСТЕЙ

У роботі розглянуто підхід до кластеризації, що базується на теорії можливостей. Такий підхід дозволяє врахувати неоднозначність вибору параметрів алгоритму кластеризації. На цій основі запропоновано методи оцінювання можливості та необхідності низки тверджень щодо точок та нечітких кластерів.

Ключові слова: множина даних, кластеризація, теорія можливостей.

Задача ефективного розбиття множини даних на класи еквівалентності за допомогою алгоритмів кластеризації має велике прикладне значення. При розв'язанні такої задачі часто виникає проблема адекватного вибору параметрів алгоритму кластеризації. При невдалому виборі цих параметрів розбиття не буде відповідати вимогам експерименту, проте навіть при вдалому їх виборі часто виникають питання щодо коректного віднесення певних підмножин множини даних до того чи іншого кластера. Пропонований в роботі підхід до кластеризації, що базується на теорії можливостей, дозволяє врахувати неоднозначність та нечіткість вибору параметрів алгоритму кластеризації.

1. Нечіткі розбиття. Кластери

Нехай $(X, 2^X, P)$ – деякий можливісний простір [1], D – певна скінчена множина даних. Позначимо як $\mathcal{D}$, де $\mathcal{D} \subseteq 2^D$, множину допустимих розбиттів D на класи еквівалентності, тобто кожне $\mathcal{A} \in \mathcal{D}$ тут таке, що $\bigcup_{i=1}^{n} A_i = D$, $A_i \in \mathcal{A}$, $|\mathcal{A}| = n$, $A_i \cap A_j = \varnothing$ для всіх $i,j \in \{1,\ldots,n\} : 1 \le i < j \le n$.

Нечітким розбиттям D на класи еквівалентності (скор. нечітким розбиттям) назвемо нечітку множину на D [2] – відображення вигляду $R : X \to \mathcal{D}$.