АВТОМАТИЧНЕ ВИЛУЧЕННЯ ДІЄСЛІВНИХ КОЛОКАЦІЙ АНГЛІЙСЬКОЇ МОВИ

Чалюк Г.Є.
Національний технічний університет
«Харківський політехнічний інститут»
м. Харків, вул. Пушкінська, 79/2, тел. 707–63–60,
e-mail: anuntus@mail.ru

Проблема вивчення синтагматичною сполучуваності та стійкості сполучень слів є однією з ключових в лінгвістиці. Існуюча література та словники не завжди повно і послідовно відображають інформацію про сполучувані переваги лексем та стійкі словосполучення. Актуальність теми обумовлена тим, що отримання нових даних про сполучуваність, розробку нових методів її вивчення повинні сприяти розвитку лексикографії, синтаксису, семантики. Робота спрямована на опис та експериментальну верифікацію лінгвістичних і статистичних прийомів виявлення колокацій в корпусах текстів на матеріалі англійської мови. Статистичний метод вилучення колокацій отримав широке поширення в корпусній лінгвістиці. Найпростішим способом виявити колокації у тексті є зіставлення переліку частотності слів, який з'являться зліва або справа ключового слова всередині даного проміжку. Розмір такого проміжку зазвичай складає 5 слів зліва або справа від ключового слова. Статистичні показники асоціацій отримали вельми широке розповсюдження у сучасних лінгвістичних дослідженнях. Ці показники базуються на частотній сполучуваності пар слів й сполучуваності кожного елемента, який можна вираховувати у межах певного проміжку.

У нашому дослідженні в якості міри асоціацій ми використовуємо коефіцієнт взаємної інформації (MI), який можна розуміти як коефіцієнт синтагматичної сили між елементами колокації.

MI порівнює залежні контекстно-пов'язані частоти з незалежними, якщо б слова у тексті з'являлись абсолютно випадково.

Коефіцієнт взаємної інформації для біграми може бути обчислений за такою формулою:

$$MI(n,c) = log_2 \frac{f(n,c) * N}{f(n) * f(c)}$$

де MI – коефіцієнт взаємної інформації;
n – ключове слово;
$f(n,c)$ – частота зустрічальності ключового слова n у парі з колокатом c;
$f(n), f(c)$ – абсолютні (незалежні) частоти ключового слова n і колоката c в корпусі;
N — загальне число словоформ в корпусі.

Проте, необхідно зауважити врахувати той факт, що слова синтаксично пов'язані, вони не зустрічаються у тексті випадково. Отже, вилучення колокації потребує не тільки статистичного методу, а й підходу, що базується на синтаксисі, який бере до уваги морфологічні й синтаксичні властивості слів у корпусі.

Застосування описуваних методів для отримання інформації про лексичну і синтаксичну сполучуваність на базі великих корпусів текстів вже сьогодні служить основою для створення словників і граматик нового типу. В даний час в сучасній лінгвістиці незамінним інструментом і одночасно матеріалом для лінгвістичних досліджень і вирішення прикладних завдань стали корпуси текстів. Об'єктом нашого дослідження виступає явище синтагматичною сполучуваності в англійській мові. Предмет дослідження - статистично стійкі поєднання (колокації), що відповідають певним лексико-синтаксичним моделям.

Планується провести ряд експериментів, метою яких є порівняння різних підходів до автоматичного вилучення колокацій дієслово-іменник. Основними темами розглядання є вплив розміру проміжку й POS-tagging, у той час як розширення розміру проміжку має неоднозначний вплив. З одного боку, це дозволяє вилучати віддалені словосполучення, але, з іншого, це призводить до помилкових сполучень, що в свою чергу призводить до розгляду використання підходу, що базується на синтаксисі для вилучення колокації дієслова та іменника.

В якості інструмента для нашого дослідження обрано IntelliText – систему, розроблену Центром для Дослідження Перекладу Університету Лідса. IntelliText надає достатню можливість для лінгвістичних досліджень і містить типові корпуси, включаючи морфологічно анотовані корпуси англійської мови.

Список литературы

1. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие / Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В. - М.: МИЭМ, 2011. - 272 с.

2. Lewis, M. Teaching Collocation: Further Development in the Lexical Approach, Hove: Language Teaching Publications, 2000.

3. Biber D., Conrad S., Reppen R. Corpus Linguistics. Investigating language structure and use. Cambridge University Press, 1998

4. Герд А.С. РНК и академическая лексикография // Труды международной конференции «Корпусная лингвистика – 2006». – СПб.: Изд-во С.-Петерб. ун-та; Изд-во РХГА, 2006. – С. 88-91.

5. Cowie, A. The Treatment of Collocations and Idioms in Learner's Dictionaries. Applied Linguistics, 1981. – 2(3). – P. 223–235.