Методы автоматического выявления устойчивых словосочетаний в научных текстах

Рожков Никита Олегович

Студент

Факультет ВМК МГУ имени М. В. Ломоносова, Москва, Россия

E-mail: ronikita98.07@gmail.com

Научный руководитель *— Ефремова Наталья Эрнестовна*

В настоящее время во многих задачах, связанных с автоматической обработкой текстов, используются различные словари устойчивых словосочетаний. Из-за быстрого развития языка и ввиду большой сложности ручного формирования подобных словарей задачи их автоматического составления и последующего обновления не теряют своей актуальности.

Для выявления устойчивых словосочетаний из текстов используют два подхода: лингвистический и статистический. Основная идея лингвистического подхода заключается в том, что потенциальными единицами словаря являются словосочетания, имеющие определенный вид, к примеру, такие, которые удовлетворяют синтаксическому образцу *«Существительное» + «Прилагательное».*

В статистических методах считается, что потенциальные единицы словаря – это словосочетания, имеющие определенное значение той или иной меры ассоциации. В общем случае, меры учитывают частотность и совместную встречаемость слов, их значения вычисляются на достаточно большой коллекции текстов. Отметим, что среди мер можно выделить две подгруппы:

- не использующие контрастную коллекцию (*MI, MI3, T-score, Log-Dice* и др.);
- использующие контрастную коллекцию (*Weirdness, Relevance, Contrastive Weight* и пр.).

Как правило, при выявлении устойчивых словосочетаний сложно опираться на значение какой-то одной меры, поскольку разные меры по-разному ранжируют получаемые словосочетания. Например, мера *MI* сильно завышает значение словосочетаний с редкими словами (в частности, со словами с опечатками). Поэтому сейчас наиболее часто используют комбинации значений меры, которые подбирают исходя из рассматриваемой задачи [1].

Помимо комбинирования можно использовать значения мер как признаки для машинного обучения [2]. При этом, в качестве признаков также могут выступать типичные для устойчивых словосочетаний синтаксические образцы.

В рамках данной работы рассматриваются задача выявления устойчивых словосочетаний русскоязычной научной прозы, для которых не существует актуальных словарей (последний из словарей был выпущен в 70-е годы прошлого века).

Для решения рассматриваемой задачи был составлен список известных общенаучных устойчивых словосочетаний (с опорой на имеющиеся словари) и собрана коллекция научных текстов, включающая статьи, учебники, энциклопедии и т. п. В качестве контрастной коллекции были рассмотрены произведения художественной литературы и новостные статьи. Для словосочетаний из словаря были вычислены значения 10 наиболее часто используемых мер и предложены несколько способов их комбинирования; один из способов основан на среднем ранге мер*MI3, T-score, Minimum-sensitivity, Log-Dice* и *MI*. Также были выявлены признаки, которые будут использованы в методах машинного обучения.

Литература

1. Дорофеева 2019 – Дорофеева А. А. Методы оценки устойчивости словосочетаний на русском языке. Выпускная квалификационная работа. МГУ им. М.В. Ломоносова. 2019
2. Нокель 2015 – Нокель М. А. Методы улучшения вероятностных тематических моделей текстовых коллекций на основе лексико-терминологической информации. Диссертация на соискание учёной степени кандидата физико-математических наук. МГУ им. М.В. Ломоносова. 2015.