Прикладная и математическая лингвистика

Динамическая тематическая модель корпуса новостных текстов на русском языке

Яна Викторовна Давыдович
Докладчик
Санкт-Петербургский государственный университет

Спорные вопросы автоматического морфологического анализа русских текстов на примере «Pymorphy2»

Ангелина Александровна Коваль
Докладчик
Санкт-Петербургский государственный университет

Синтаксическая деривация как способ повышения эффективности алгоритма выделения ключевых слов, основанного на тематической атрибуции

Анна Александровна Токарева
Докладчик
Санкт-Петербургский государственный университет

«Ключевые слова» русской прозы начала XX в.

Анна Андреевна Гусева
Докладчик

ению русского языка
и синхронические
аспекты фонетики, лексики, грамматики, стилистики, диалектологии русского языка, вопросы
лингвофольклористки и преподавания русского языка как иностранного.

Динамическая тематическая модель корпуса новостных текстов на русском языке

пригодны для работы с корпусами текстов, снабженными временными маркерами. Для решения данной проблемы был разработан алгоритм динамического тематического моделирования DTM, реализованный в библиотеке gensim для Python (https://radimrehurek.com/gensim/models/ldaseqmodel.html). Отличительной особенностью подходов динамического тематического моделирования является отслеживание семантических и сочетаемостных изменений в словаре корпусов текстов и, тем самым, регистрация изменений в тематике текстов с течением времени.

Целью нашего исследования является автоматическое исследование тематики новостного потока в заданных хронологических рамках. В связи с этим, методологическую основу исследования составляет процедура динамического тематического моделирования.

Материалом исследования является корпус сообщений новостного портала «Коммерсантъ» (https://www.kommersant.ru/). Нами было создано программное обеспечение для автоматического формирования корпуса на основе библиотеки BeautifulSoup для Python (https://pypi.org/project/beautifulsoup4/). При парсинге данных новостного портала мы описались на его внутреннюю рубрикацию. В качестве анализируемых сегментов были выбраны новости следующих: мир, общество, политика, культура, спорт. Анализировались новости за 2008, 2014 и 2019 гг. Каждый из данных промежутков времени характеризуется специфическими экстралингвистическими реалиями, которые были отражены в СМИ (к примеру, олимпиада в Сочи 2014 г.).

Данное исследование доказывает состоятельность процедуры динамического моделирования тем на основе алгоритма неотрицательной матричной факторизации (NMF) (https://radimrehurek.com/gensim/models/nmf.html). Результаты исследования связаны с проверкой гипотезы о том, что событийное разнообразие в различных сферах жизни общества и всплеск внимания СМИ к этим событиям даст наиболее показательные результаты использования DTM в определении тем и сопоставлении тематического состава новостей за различные промежутки времени.

Динамическая тематическая модель корпуса новостных текстов на русском языке

Яна Викторовна Давыдович
Докладчик
студент 2 курса
Санкт-Петербургский государственный университет

Ключевые слова, аннотация

В данной работе описываются результаты автоматического построения динамической тематической модели (DTM) для корпуса новостных сообщений на русском языке. Корпус создан автором на основе данных СМИ и имеет внутреннюю рубрикацию (новости общества, культуры и т. д.). Выбранный нами подход позволяет проследить хронологические изменения в тематике текстов и их связь с событиями в мире и обществе. Основной задачей исследования являлась содержательная интерпретация тем, выделенных с помощью DTM, и сравнение результатов классической модели LDA и динамической модели DTM для корпуса новостей.

Тезисы

Мониторинг новостного потока уже долгое время является одной из приоритетных прикладных задач компьютерной лингвистики. Актуальность исследований возрастает с расширением числа электронных СМИ и с усилением влияния социальных медиа на общественное сознание. Сейчас в извлечении информации из новостных текстов применяются новые методы и модели, среди которых на первый план выходят процедуры тематического моделирования, позволяющие быстро определить соотношение текстов по тематике и произвести нечеткую рубрикацию новостей.

Стандартные подходы к тематическому моделированию (LSA, pLSA, LDA и т.д.) не позволяют учитывать хронологический порядок появления документов в корпусе, поэтому они не пригодны для работы с корпусами текстов, снабженными временными маркерами. Для

Санкт-Петербургский государственный университет

23 ОТКРЫТАЯ КОНФЕРЕНЦИЯ СТУДЕНТОВ-ФИЛОЛОГОВ

НАСТОЯЩАЯ СПРАВКА ПОДТВЕРЖДАЕТ, ЧТО

ДАВЫДОВИЧ ЯНА ВИКТОРОВНА

ВЫСТУПИЛА С ДОКЛАДОМ

ДИНАМИЧЕСКАЯ ТЕМАТИЧЕСКАЯ МОДЕЛЬ КОРПУСА НОВОСТНЫХ ТЕКСТОВ НА РУССКОМ ЯЗЫКЕ

В РАМКАХ

XXIII ОТКРЫТОЙ КОНФЕРЕНЦИИ СТУДЕНТОВ-ФИЛОЛОГОВ САНКТ-ПЕТЕРБУРГСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА
26–31 ОКТЯБРЯ 2020 Г.

(ПЕРВОНАЧАЛЬНЫЕ СРОКИ ПРОВЕДЕНИЯ КОНФЕРЕНЦИИ – 13–17 АПРЕЛЯ 2020 Г.)

[redacted]

СОПРЕДСЕДАТЕЛЬ ПРОГРАММНОГО КОМИТЕТА — О. В. БЛИНОВА

СОПРЕДСЕДАТЕЛЬ ПРОГРАММНОГО КОМИТЕТА — В. А. ТУРЧАНЕНКО