Лонин В.А.  
Научный руководитель: Кандидат филологических наук,   
Доцент Учебно-научного центра лингвистической типологии Коротаев Н.А.

Автоматическое определение года создания текста: анализ языковых характеристик на материале Живого Журнала

Работа посвящена анализу лингвистических характеристик текстов, влияющих на автоматическое определение года его создания на материале Живого Журнала. Мы определяем прогнозируемое время создания / написания текстов, следовательно, определяем и язык этого года – какая лексика использовалась, какие речевые обороты, порядок слов, грамматические и лексические приемы, англицизмы, неологизмы и так далее.

Иными словами, мы занимаемся дистрибутивной семантикой, анализируем изменение дистрибутивного распределения лексических единиц в тексте и степени их семантической близости на временном отрезке. В результате анализа мы можем получить представление, как язык изменялся на протяжении времени.

Основное предположение, лежащее в основе этой работы, заключается в том, что естественный язык демонстрирует тенденцию к изменению частот слов с течением времени. Новые слова постоянно входят в обиход, в то время как другие слова выходят из употребления либо после кратковременного увлечения, либо когда они становятся устаревшими или архаичными. Текущие события, популярные проблемы и темы также влияют на выбор слов авторами, а также на период времени, когда они создают документы( Angelo Dalli, Yorick Wilks).

Одним из первых исследований по моделированию временной информации для автоматического датирования документов является работа де Йонга и др. (2005). В этих экспериментах авторы использовали языковые модели unigram для классификации голландских текстов, охватывающих период с января 1999 по февраль 2005 года, с использованием нормализованного отношения логарифмического правдоподобия (NLR) (Kraaij, 2004) (Vlad Niculae , Marcos Zampieri , Liviu P. Dinu , Alina Maria Ciobanu, 2014). Нашем объектом исследований будет материал, взятый с Живого Журнала в период в 2001 по 2020 года.

Наша цель - исследовать изменения в дистрибутивном распределении лексических единиц в текстах в зависимости от времени их создания.

Наши задачи:

1. Собрать необходимый материал с разметкой корпуса по годам.
2. Создать алгоритм, прогнозирующий год создания.
3. Обучить его анализировать тексты.
4. Провести ручной анализ, составить частотные словари.
5. Получить необходимые результаты, сделать выводы.

Есть две гипотезы:

Основная: Изменения в дистрибуции есть, влияет, в основном, тематика.

Альтернативная: Тематика не влияет, эволюции нет.

Для реализации этих задач и проверки гипотез понадобится мегакорпус. В данном случае, материал – тексты, взятые из ЖЖ с автоматической разметкой текста по годам (2001 – 2020), ВК (2014-2017) для дополнительных экспериментов.

Мы разделим его на обучающую выборку, проверочную и тестовую. На обучающей выборке задействуем определённое количество классификаторов. Целевым признаком будет служить год создания текстов, а особенностями – грамматическая и лексическая структура предложений, семантическая близость в качестве признаков будем использовать вектора слов, составленные с помощью таких методов, как BoW и TF-IDF. Чтобы качество метрик было выше, гиперпараметры обучения классификаторов будут подвергаться ряду изменений, а также будут проведены эксперименты с тематическим моделированием – так мы определим максимально эффективную настройку моделей.

Классификатор выявит закономерности между особенностями и целевым признаком и научится определять год, в зависимости от лингвистических характеристик. На проверочной выборке мы убедимся в его работоспособности, а на тестовой – в правильности его решений.

Немаловажную роль будет играть и ручной анализ – подсчёт статистики частотности слов – так мы получим более детальный анализ, подкрепляющий, или опровергающий нулевую гипотезу.

Библиография:

Angelo Dalli, Yorick Wilks, Automatic Dating of Documents and Temporal Text Classification

Vlad Niculae , Marcos Zampieri , Liviu P. Dinu , Alina Maria Ciobanu, 2014, Temporal Text Ranking and Automatic Dating of Texts

Anne Garcia-Fernandez , Anne-Laure Ligozat , Marco Dinarelli, and Delphine Bernhard, 2011, When was it Written, Automatically Determining Publication Dates

H. Abe and S. Tsumoto. 2010. Text categorization with considering temporal patterns of term usages. In Proceedings of ICDM Workshops, pages 800– 807. IEEE.