МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ ИМЕНИ Н. Э. БАУМАНА

Факультет информатики и систем управления

Кафедра теоретической информатики и компьютерных технологий

Курсовой проект

по курсу «Конструирование компиляторов»

«Препроцессор синаксического сахара для языка Scheme»

Выполнил:
студент ИУ9-101
Выборнов А. И.

Руководитель:
Дубанов А. В.

Москва 2015

Содержание

Введение

1. Обзор литературы

1.1. Linking Tweets to News: A Framework to Enrich Short Text Data in Social Media

1.1.1. Перевод аннотации

Многие современные методы обработки естественного языка (NLP[1]) хорошо работают, используя большой массив текста в качестве входных данных. Однако они становятся неэффективными, когда применяются на коротких текстах, таких как твиты. Чтобы преодолеть эту проблему, мы хотим найти соответствующий данному твиту новостной документ, для большей эффективности NLP задач. Это требует хорошего моделирования и понимания семантики коротких текстовых твитов.

Вклад статьи двойной: 1. мы представим задачу связывания твитов с новостями, а также набор пар твит-новости, из этого могут извлечь выгоду многие NLP задачи; 2. в отличие от предыдущих исследований, которые фокусируются на лексических особенностях в коротких текстов (информация о связи текст-слово), мы предлагаем граф, основанный на модели скрытой переменной, которая моделирует корреляцию между короткими текстами (информация о связи текст-текст). Это обосновано наблюдением: твит обычно покрывает только один аспект события. Мы покажем, что с помощью особенных признаков твита (хэштегов) и особых признаков новостей (именнованные сущности[2]) а также временны́х ограничений, мы можем получить взаимосвязь текст-текст, и, таким образом, дополнить семантическую картину короткого текста. Наши эксперименты показывают значительное преимущество нашей новой модели над baseline[3] для трёх методов оценки.

1.1.2. Краткое изложение статьи

Современные NLP подходы на коротких текстах плохо работают или не работают вообще. Чтобы это преодолеть предлагается к твитам привязывать соответствующие новости.

Для train выкачали твиты за 18 дней, которые имели ссылки на CNN или NYT, опубликованные в этот период.

[1]Natural Language Processing

[2]Какой-то кривой перевод, найдо найти получше. In data mining, a named entity is a phrase that clearly identifies one item from a set of other items that have similar attributes.

[3]Как перевести?

Модели со скрытой переменной хорошо подходят для отображения коротких сообщений в малоразмерный вектор.

Построили модель со скрытой переменной, чтобы применять её и к твитам и к новостям. Протестили модель на наборе данных из небольших сообщений: с большим запасом превзошла и LSA и LDA.

1.1.3. Используемые технологии

1.1.4. Основная идея

1.1.5. Полученные результаты

1.2. Linking Online News and Social Media

1.2.1. Перевод аннотации

Многое из того, что обсуждается в социальных медиа вдохновлено событиями, описанными в новостях и, наоборот, социальные медиа предоставляют механизм, позволяющий влиять на новостные события. Мы обращаемся к следующей связывающей задаче: по новости, найти в социальных сетях высказывания, которые неявно на неё ссылаются. Мы используем трехступенчатый подход: сначала получаем несколько моделей запросов по исходной статье, модели используются для получения высказываний из индекса целевого социального медиа, в результате получаем несколько ранжированных списков, которые объединяются с использованием техники слияния данных. Модели запроса создаются на основе структуры исходной статьи и явно связанных высказываний из социальных медиа, в которых обсуждается исходная статья. Для борьбы с дрейфом запроса[1] в результате большого объема текста, либо в самой исходной новости, либо в явно связанных высказываниях в социальных медиа, мы предлагаем основанный на графике[2] метод для выбора отличительных условий.

Для нашей экспериментальной оценки, мы используем данные из Twitter, Digg, Delicious[3], the New York Times Community, Wikipedia и блогосферы, для порождения моделей запросов. Мы покажем, что другие модели запросов, основанные на различных источниках данных, обеспечивают дополнительную информацию и влияют на получение различных высказываний из социальные медиа по нашему целевому индексу. Как следствие, методы слияния данных приводят к значительному

[1] Порождение менее подходящего запроса.

[2] Или всё же графах?

[3] Веб-сайт, бесплатно дающий зарегистрированным пользователям услугу хранения и публикации закладок на страницы Всемирной сети.

повышению производительности в сравнении с индивидуальными подходами. Показано, что основанный на графике метод выделения условий помог улучшить как эффективность, так и продуктивность.

1.2.2. Используемые технологии

1.2.3. Основная идея

1.2.4. Полученные результаты

Ещё одна статья от Лукашевич

1.3. Bridging Vocabularies to Link Tweets and News

Ещё один подход

1.4. TwitterStand: News in Tweets

2. Дополнительная литература

2.1. Gibberish, Assistant, or Master? Using Tweets Linking to News for Extractive Single-Document Summarization

2.2. Detecting Event-Related Links and Sentiments from Social Media Texts

2.3. Определение тематической направленности текстового содержимого микроблогов

2.4. Разработка сервиса извлечения мнений

3. Полезные ссылки

- Твиттер NYT: https://twitter.com/nytimes (20млн подписчиков, 200тыс твитов)
- Крайне отстойная статья на тему: http://cyberleninka.ru/article/n/issledovanie-otklika-polzovateley-twitter-na-novosti-iz-smi
- Идея для формирования train: http://techcrunch.com/2013/08/19/twitter-related-headlines/

Список литературы

[1] W. Guo, H. Li, H. Ji, and M. T. Diab. Linking tweets to news: A framework to enrich short text data in social media. - ACL, pages 239–249, 2013.

[2] T. Hoang-Vu, A. Bessa, L. Barbosa and J. Freire. Bridging Vocabularies to Link Tweets and News. - International Workshop on the Web and Databases (WebDB 2014), Snowbird, Utah, US, 2014.

[3] J. Sankaranarayanan, H. Samet, B. Teitler, M. Lieberman, J. Sperling. TwitterStand: news in tweets. - 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, 2009, Seattle, Washington.

[4] Matthew Flatt, Robert Bruce Findler. The Racket Guide. Racket Documentation: URL: http://docs.racket-lang.org/guide/index.html.

[5] R. Kelsey, W. Clinger, J. Rees. Revised^5 Report on the Algorithmic Language Scheme. Higher-Order and Symbolic Computation, Vol. 11, No. 1, August, 1998

[6] Simon Marlow. Haskell 2010 Language Report, 2010.

[7] Terence Parr. The Definitive ANTLR 4 Reference. 2013.

[8] Ахо, Альфред В., Лам, Моника С., Сети, Рави, Ульман, Джеффри Д. Компиляторы: принципы, технологии и инструментарий, 2-е изд.: Пер. с англ. – М.: ООО «И.Д.Вильямс», 2008 – 1184 с.

[9] Макконнелл С., Совершенный код. Мастер-класс: Пер. с англ. — М.: Издательство «Русская редакция», 2014 — 896 стр.