1. Обработка Естественного Языка

Наш метод извлечения кризисных событий подразумевает необходимость в комплексном препроцессинге текстов на естественном языке. В частности, мы проводим базовый лингвистический анализ, извлечение именованных сущностей, распознавание временных сущностей и классификацию текстов по тому, связаны ли они с чрезвычайными ситуациями.
Данные, полученные в результате препроцессинга используются для следующих задач: сфокусированного сбора данных, дополнения информации о событиях и создании модулярности для тематического моделирования.

1.1. Базовый Лингвистический Анализ

Этап базового лингвистического анализа состоит из токенизации, разделения предложений, определения частей речи, лемматизации и синтаксического разбора. Данные операции проведены с помощью библиотеки $IsaNLP$[1], объединяющей в себе различные компоненты, предназначенные для Русского и Английского языков. В нашей работе мы проводим эксперименты только на корпусе текстов на Английском языке. Токенизация, разделение предложений, определение частей речи и лемматизация проведены с помощью библиотеки NLTK [4], а для синтаксического разбора мы используем SyntaxNet McParseface [1].

1.2. Извлечение именованных сущностей

Для дальнейшего использования мы извлекаем из текстов следующие типы сущностей: имена людей, названия организаций, географические локации и названия кораблей. Фреймворк Polyglot используется для базового извлечения сущностей. Он использует модель, обученную на корпусе текстов Wikipedia и может использоваться для текстов на 40 языках. Однако, стоит заметить, что данный подход не применим для извлечения локаций из-за недостаточной полноты извлечения. Это делает невозможным качественную фильтрацию текстовых потоков, а так же усложняет тематическое моделирование. Причиной тому служит специфика текстов из Википедии, а именно отсутствие в них некоторых особых локаций. Polyglot так же не предоставляет функционал для нормализации извлеченных локаций.
В предыдущей работе, для того, чтобы увеличить полноту извлечения и получить возможность нормализации текстового описания локации в вид географических координат, мы создали модуль, основанный на правилах и словарях. В его основе лежит список сущностей (gazetteer) из данных Geonames, а так же несколько правил для фильтрации, основанные на частях речи извлеченных токенов. Geonames так же предоставляет географические координаты для каждой локации в своей базе, что

делает возможным нормализацию.
Для извлечения и нормализации временных сущностей мы скомбинировали два инструмента: spaCy(фреймворк для NLP, основанный на нейронных сетях) и datetimeparser.
Как и в предыдущей работе, для извлечения названий кораблей мы применили гибридный подход. Во-первых, основываясь на базе данных названий кораблей мы создали gazetteer с высокой полнотой извлекаемых сущностей, но низкой точностью, из-за того, что в названиях большинства кораблей встречаются достаточно общие слова. Для борьбы с этим явлением мы натренировали нейронную сеть, основанную на архитектуре C-LSTM. Сеть отфильтровывает имена, полученные из gazetter-a и позволяет значительно увеличить точность и среднее значение F_1-меры.

1.3. Обнаружение кризисных сообщений

Так же, как и в предыдущей работе, мы использовали гибридный подход из gazetteer-a на базе CrisisLex[17] лексикона и нейросети, основанной на архитектуре C-LSTM. Так же, как и в предыдущем случае, gazetteer генерирует много ложноположительных результатов, которые отфильтровываются нейронной сетью. Нами был собран датасет сообщений из Twitter, использованных для обучения. В данной работе мы улучшили результаты путем использования большего датасета сообщений из CrisisLex и путем рассмотрения:

- Различных видов эмбеддингов: fastText [11] (натренированный на наших данных/на датасете из Википедии), GloVe[12] (на CommonCrawl/Twiiter сообщений), Word2Vec[3], InferSent [2].
- Различных моделей: Logistic Regression и Random Forest (из библиотеки sklearn), Gradient boosted decision trees (реализованный в LightGBM[7]), Fully-Connected Network (FCN), Convolutional network (CNN), а также C-LSTM из прошлой работы.

Для использования Logistic Regression, Random Forest, Gradient Boosting и FCN мы усреднили вектора эмбеддингов что бы получить необходимые признаки. Для CNN и C-LSTM вектора эмбеддингов использовались без усреднения. Признаки, полученные с помощью алгоритма Infersent не использовались для CNN и C-LSTM моделей, т.к этот алгоритм проецирует весь текст в один вектор признаков, тогда как упомянутые модели требуют векторы для каждого из слов в сообщении.
Полносвязная нейронная сеть (FCN) представляет собой простой двухслойный перцептрон со слоем dropout посередине. Функция активации первого слоя - ReLU, второго - softmax. Сверточная сеть (CNN) основана на архитектуре, предложенной в [9]. Сначала, предложения выравниваются по длине и подаются на сверточные слои и далее на слой maxpooling, для уменьшения размерности. Полученные вектора при-

знаков объединяются в один и подаются на линейный слой для генерации предсказаний. Функции активации у сверточного и линейного слоя - ReLU и softmax соответственно.Архитектура сети C-LSTM основана на одном сверточном слое с функцией активации ReLU и слоем maxpooling после. Далее идет реккурентный слой LSTM. Финальные предсказания получаются двумя линейными слоями с активациями Tanh и softmax. Все нейронные сети реализованы на фреймворке PyTorch[10].

Список литературы

[1] Daniel Andor, Chris Alberti, David Weiss, Aliaksei Severyn, Alessandro Presta, Kuzman Ganchev, Slav Petrov, and Michael Collins. Globally normalized transition-based neural networks. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), volume 1, pages 2442–2452, 2016.

[2] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, A. Bordes, Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. arXiv preprint arXiv:1705.02364.

[3] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient Estimation of Word Representations in Vector Space

[4] Bird, Steven, Edward Loper and Ewan Klein (2009), Natural Language Processing with Python. O'Reilly Media Inc.

[5] L. Breiman, "Random Forests", Machine Learning, 45(1), 5-32, 2001.

[6] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification Journal of Machine Learning Research 9(2008), 1871-1874.

[7] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. "LightGBM: A Highly Efficient Gradient Boosting Decision Tree". In Advances in Neural Information Processing Systems (NIPS), pp. 3149-3157. 2017.

[8] Vinod Nair and Geoffrey Hinton. Rectified Linear Units Improve Restricted Boltzmann Machines. ICML. 2010

[9] Yoon Kim. Convolutional Neural Networks for Sentence Classification. http://arxiv.org/abs/1408.5882

[10] Paszke, Adam and Gross, Sam and Chintala, Soumith and Chanan, Gregory and Yang, Edward and DeVito, Zachary and Lin, Zeming and Desmaison, Alban and Antiga, Luca and Lerer, Adam. Automatic differentiation in PyTorch. NIPS-W. 2017

[11] P. Bojanowski*, E. Grave*, A. Joulin, T. Mikolov, Enriching Word Vectors with Subword Information. Transactions of the Association for Computational Linguistics. 2017. Vol. 5. pp 135-146.

[12] Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. GloVe: Global Vectors for Word Representation. 2015.

[13] Roy Chowdhury S, Purohit H, Imran M. D-sieve: a novel data processing engine for efficient handling of crises-related social messages. InProceedings of the 24th International Conference on World Wide Web 2015 May 18 (pp. 1227-1232). ACM.

[14] Zhang S, Vucetic S. Semi-supervised discovery of informative tweets during the emerging disasters. arXiv preprint arXiv:1610.03750. 2016 Oct 12.

[15] Li H, Caragea D, Caragea C, Herndon N. Disaster response aided by tweet classification with a domain adaptation approach. Journal of Contingencies and Crisis Management. 2018 Mar;26(1):16-27.

[16] Zhou C, Sun C, Liu Z, Lau F. A C-LSTM neural network for text classification. arXiv preprint arXiv:1511.08630. 2015 Nov 27.

[17] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. CrisisLex: A lexicon for collecting and filtering microblogged communications in crises. In Proceedings of ICWSM, 2014.