Отчёт о выполнении эксперимента «Поиск географических названий в тексте»

Выполнил: Плетнев Никита, 574

1. Происхождение задачи

Задача взята с конкурса SemEval 2019 и является первой частью масштабной программы. Требуется разработать алгоритм, позволяющий находить в тексте научной статьи топонимы. Предполагается, что полученные топонимы будут разыскиваться в базе данных с целью определения набора мест (то есть, меток на карте), к которым имеет отношение данная статья.

2. Постановка задачи и данные

Дан текст — научная статья небольшого размера в файле txt. Требуется определить в нём все топонимы. Для проверки к статье приложен файл ann, в котором перечислены все входящие в текст топонимы и соответствующие им метки.

Данная задача эквивалентна следующей задаче бинарной классификации: дан текст — последовательность слов. Каждое слово принадлежит одному из двух классов. Первый класс (обозначим 1) — это слова, которые являются топонимами или их частями (топонимы бывают и составными). Второй класс (0) — слова, не имеющие отношения к географическим названиям.

Требуется построить функцию

$$a : T \times \mathrm{N} \longrightarrow \{0, 1\},$$

где T — множество текстов, а N — множество номеров слов в тексте, которая наиболее точно приближала бы функцию

$$y(t, i) = \begin{cases} 1, t[i] \text{ — топоним или часть топонима;} \\ 0, \text{ иначе.} \end{cases}$$

Набор данных содержит 55 размеченных текстов. Они перемешиваются в произвольном порядке, после чего часть из них составляет обучающую выборку, а все остальные — тестовую.

3. Методы оценивания результатов

Как и для всякой задачи бинарной классификации, здесь существуют две метрики точности — precision и recall. Расшифровка обозначений приведена в таблице (столбец — предсказанный класс, строка — истинный класс; на пересечении — количество измерений, соответствующих данным значениям):

	1	0
1	TP	FN
0	FP	TN

Тогда

$$precision = \frac{TP}{TP + FP}, \quad recall = \frac{TP}{TP + FN}.$$

Максимизация каждой из этих величин приводит к уменьшению другой, поэтому используется их среднее гармоническое:

$$f1 - score = \frac{2precision \times recall}{precision + recall}.$$

4. Обзор литературы

По темам, близким к рассматриваемой задаче, были обнаружены всего две публикации: статья What's missing in geographical parsing? коллектива Milan Gritta, Mohammad Taher Pilehvar, Nut Limsopatham, Nigel Collier и тезисы Jochen Lothar Leidner, посвященные проблеме Toponym Resolution in Text.

Авторы решают задачи второй и третьей части конкурса: разделение соответствующих названиям географических объектов и сопоставление им конкретных точек на карте.

Однако в обеих работах считается известным словарь, в котором лежат все топонимы. Это значительно упрощает обнаружение их в тексте. Здесь же будет решаться другая задача: найти топонимы по контексту, не используя словарь.

5. Применяемые методы

Для решения задачи применялись такие методы, как наивный подход (определение топонимов на основании очевидных и легко интерпретируемых признаков), многоклассовая классификация и контекстный анализ.

Сводная таблица результатов:

	precision	recall	f1-score
Naive	25%	30%	30%
Multy	?	?	?
Context	90%	86%	88%

Рассмотрим методы подробнее.

6. Наивный подход

Очевидный и неинтересный способ решения задачи; результаты у него соответствующие. Предполагается, что если слову предшествует предлог места (into, in, at, from), при этом оно является именем существительным и начинается с заглавной буквы, то оно с некоторой вероятностью, которой нельзя пренебречь, является топонимом. Также встречается ситуация, когда географическое название обособлено с двух сторон запятыми; с заглавной буквы оно начинается всегда.

Поскольку использованные тексты содержат достаточно много существительных, начинающихся с заглавной буквы и не являющихся топонимами, достичь высокой точности на этом пути невозможно.

7. Многоклассовая классификация

Текст параметризуется с помощью TfidfTransformer и HashingVectorizer. На полученном наборе векторов (текстов) решается задача множественной классификации. Каждому топониму соответствует свой класс; вопрос о принадлежности текста каждому классу решается с помощью мультиклассового классификатора OneVsRestClassifier, обучаемого на выборке текстов, для которых известно, в какие классы они входят.

Данный алгоритм очень точно обнаруживает те топонимы, которые есть в обучающей выборке (более 95%). Однако новые топонимы, которые на обучающей выборке не встречались, не определяются. Поэтому следует либо использовать обучающую выборку большего объема, либо применять рассмотренный метод в сочетании с другими.

Малое количество текстов приводит к тому, что этот метод невозможно протестировать — тестовая выборка обязательно содержит топонимы (то есть классы), которые не встречаются в обучающей выборке. Поэтому данная модель оказывается бесполезной.

8. Контекстный анализ

Идея метода: классификация производится на основании ближайшего окружения слова в тексте. Наивный подход использует частные случаи признаков такого рода. Здесь же применяются предобученные эмбединги. Вектор признаков получается путем конкатенации или суммирования эмбедингов слов в окне заданного размера, содержащем исследуемую позицию. В качестве обучающей выборки берутся 40 текстов. Остальные 15 текстов используются для контроля.

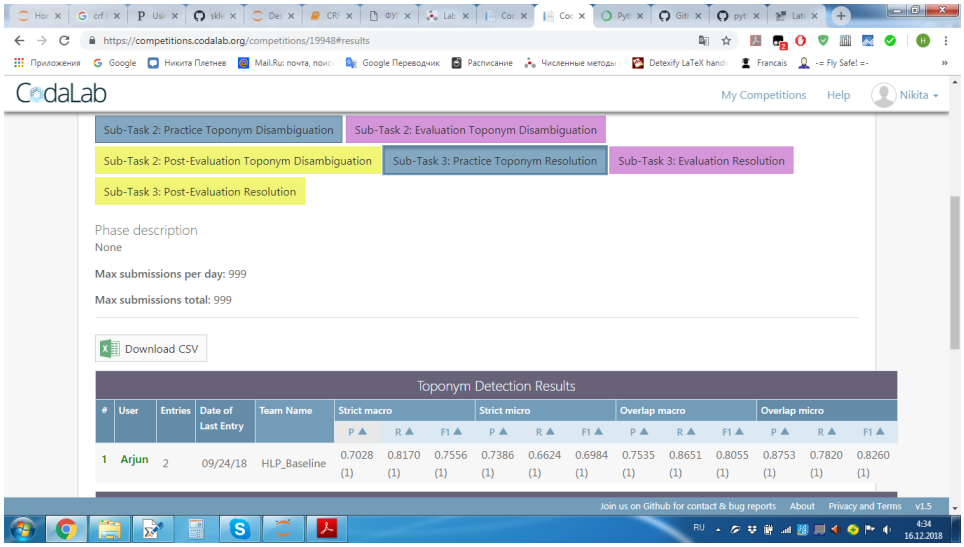
Сравнение результатов для 3 предшествующих слов и k последующих. Слева результаты для конкатенации, справа — для суммы:

k	precision	recall	f1-score
3	90%	83%	87%
1	90%	86%	88%
0	90%	83%	86%

k	precision	recall	f1-score
3	64%	42%	51%
1	70%	48%	57%
0	69%	45%	55%

9. Результаты

Соревнование официально еще не началось, и заявка на участие не одобрена. Однако контекстный анализ показал результаты лучше, чем единственная запись в таблице.



Итак, в результате эксперимента предложен метод, дающий лучшие результаты из тех, которые можно наблюдать на текущий момент. В качестве продолжения работы в данном направлении можно попробовать применить для обнаружения географических названий нейросеть CFR.