УДК 004.048

РАСПОЗНАВАНИЕ ИМЕНОВАННЫХ СУЩНОСТЕЙ С ПОМОЩЬЮ DOCCANO

Яшихина Е.В. (Национальный исследовательский университет ИТМО)
Научный руководитель – к.э.н., ординарный доцент ИДУ Кононова О.В.
(Национальный исследовательский университет ИТМО)

В данной работе приводятся виды аннотации и основные функциональные возможности инструмента Doccano. Рассмотрены подходы к решению задачи распознавания именованных сущностей и сложность ее реализации.

Введение.

Модели машинного обучения нуждаются в размеченных данных, но большинство данных, как правило, собираются в необработанном формате. Поэтому первым шагом перед построением модели будет подготовка данных – разметка корпусов текстов экспертами предметной области. Для этого стоит использовать инструмент аннотации текста с открытым исходным кодом Doccano. Он позволяет создать набор данных за несколько часов, создав проект, загрузив данные и начав аннотацию.

Основная часть.

В Doccano различают следующие типы аннотаций:

— ручная аннотация (пользователь может начать определять разметку, после назначения сущностей. нажав на кнопку метки);

— полуавтоматическая аннотация (модель машинного обучения загружается для автоматического аннотирования загруженных данных; пользователь может внести некоторые исправления в неправильные прогнозы);

— автоматическая аннотация (загруженный документ автоматически аннотируется обученным документом модели машинного обучения).

Существует три основных функции, которые можно выполнить с помощью Doccano. Каждая функция имеет свой собственный эффективный формат вывода:

— анализ настроений – это задача классификации текстов и тем по различным категориям. Поскольку текст может относиться к нескольким категориям, аннотация может быть многозначной.

— машинный перевод – одна из последовательных задач, которая позволяет выполнять несколько ответы, если может быть предоставлено более одного ответа.

— распознавание именованных сущностей (Named Entity Recognition, NER) – одна из задач маркировки последовательностей. Текст выбирается и аннотируется в соответствии с определенными сущностями. Необходимо обнаружить, какая из последовательности слов – это именованная сущность и понять, к какому классу эта именованная сущность относится.

Стандартного набора классов нет, все зависит от задачи исследования. Наибольшее распространение для широкого спектра задач получила выделение таких сущностей, как персона (Per) – имена, фамилии, отчества людей; локация (Loc) – местоположение; организация (Org) – названия организаций, компаний, объединений; разное (Misc). В эту группу входят все прочие типы сущностей, если их более тщательное разделение не требуется для целей исследования.

На первый взгляд кажется, что с именованными сущностями не должно возникнуть особых сложностей. Ведь большинство из них являются именами собственными, которые пишутся с большой буквой. Но во многих языках, таких как китайский или арабский, нет больших букв, а в некоторых — не только имена собственные пишутся с большой буквы. Например, в немецком языке с большой буквы пишутся вообще все существительные.

Также вызывает трудности обстоятельство, что именованные сущности редко состоят из одного слова и не всегда очевидны границы сущности. Поэтому различают начало,

середину и конец именованной сущности. Для обозначения первого слова используют префикс «B» (beginning), для последнего слова — «E», а для всех слов между — «I» (intermediate). Таким образом, задача сводится к пословной классификации.

Обычно для построения классификатора на большом количестве текстов с разметкой именованных сущностей тренируют нейросетевую модель. Хорошие результаты дают и классические классификаторы, работающие с предзаданным множеством признаков. Для решения задач NER также используют именованные сущности, которые уже собраны в списки, газетиры. Еще применяются системы, основанные на правилах, в них прописываются шаблонные схемы именованных сущностей. Но проблема такого подхода заключается в том, что подготовка правил требует очень много времени, а малейшее отступление от них приведет к ошибке.

Выводы.

Аннотация важна не только для создания наборов данных, но и для понимания области исследования. Данная работа дает глубокое понимание данных, возможность экспертам обсудить разногласия и лучше понять трудности в области исследования.

Работа выполнена в рамках проекта НИР Университета ИТМО № 621304 «Разработка сервиса тематической кластеризации корпуса текстов «Развитие цифрового государственного управления в Российской Федерации» на основе машинного обучения».