**Практическая домашняя работа 1**

**“Маркировка семантической роли”**

**Baseline решение**

**Описание задания:**

Задача мотивирована потребностями человека сравнивать различные объекты: разные модели мобильных телефонов, автомобилей, языки программирования, страны и т.д. Исследования NLP частично решают данную задачу сравнения объектов, однако в настоящее время существует множество возможностей для улучшения существующих сравнительных систем.

Например, система CAM (Comparative Argument Mining) получает пару объектов для сравнения и извлекает аргументы в пользу каждого из них. Аргументами в сравнительных предложениях служат предикаты (сравнительные характеристики объектов, например, проще, лучше, быстрее и т.д.) и аспекты (характеристики, по которым сравниваются объекты, например, скорость, экран, производительность и т.д.). Аспекты и предикаты извлекаются с использованием рукописных шаблонов, которые имеют низкую полноту (Recall) – не удается извлечь объекты, которые не соответствуют шаблонам, – а иногда извлекаются неправильные объекты.

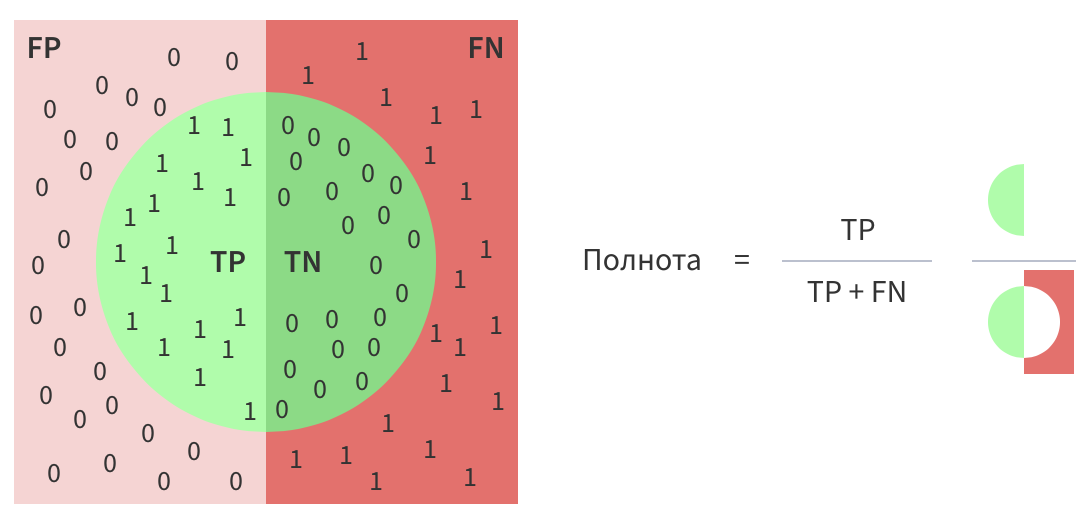


Рис. 1. Полнота (Recall) - это доля найденных объектов, принадлежащих классу относительно всех объектов этого класса в выборке. *TP* — истино-положительное решение; TN— истино-отрицательное решение; FP — ложно-положительное решение; FN — ложно-отрицательное решение.

В данном задании вам предлагается воспроизвести baseline решение извлечения аргументов (объектов, аспектов и предикатов) из предложения и оценить качество этого решения. Такая модель должна быть обучена на предложениях, где слова или фразы имеют разметку последовательности – каждому слову соответствует его тег.

**Примеры предложений**

Postgres is easier to install and maintain than Oracle.

*[Postgres OBJECT] is [easier PREDICATE] to [install ASPECT] and [maintain ASPECT]*

*than [Oracle OBJECT].*

Сущности могут состоять из нескольких слов:

Advil works better for body aches and pains than Motrin.

*[Advil OBJECT] works [better PREDICATE] for [body aches ASPECT] and [pains*

*ASPECT] than [Motrin OBJECT].*

**Формат данных**

Представленные файлы данных имеют формат CoNLL. Каждая строка содержит одно слово и его метку, разделенные табуляцией ("Word<TAB>label"), конец предложения отмечен пустой строкой. Метки представлены в формате BIO, где каждая из меток сущности ("Объект", "Аспект", "Предикат") предваряется префиксом "B-" или "I-", указывающим начало сущности (первое слово сущности) и внутреннюю часть объекта. сущность (второе и все последующие слова). Слова, которые не являются частью сущности, помечаются буквой "О".:

advil B-Object works O  
better B-Predicate for O

body B-Aspect aches I-Aspect and O  
pains B-Aspect than O

motrin B-Object

**Формулировка задачи**

Данные состоят из сравнительных предложений (т.е. предложений, содержащих сравнение двух или более объектов). В предложениях содержится следующая информация:

* + Объекты – объекты, которые сравниваются
  + Аспекты – характеристики, по которым сравниваются объекты
  + Сказуемое – слова или фразы, которые реализуют сравнение (обычно сравнительные прилагательные или наречия)

*В наборе данных используется схема BIO:*

● Первое слово сущности помечается тегом “B-<entity-type>” (начало сущности). ● Второе и последующие слова сущности помечаются тегом “I-<entity-type>” (внутри сущности).  
● Слова, которые не являются частью сущности, помечаются тегом “O” (вне сущности).

Поэтому в нашем наборе данных используются следующие метки:

● sO

● B-Object  
● I-Object  
● B-Aspect  
● I-Aspect  
● B-Predicate

● I-Predicate

*Ваша задача* – присвоить одну из таких меток каждому токену данных путем воспроизведения baseline решения, а также проверка качества решения.

**Метрики оценки качества**

Результаты будут оценены при помощи метрики F1-score:

2 ⋅ 𝑝𝑟𝑒𝑐𝑖𝑠𝑖𝑜𝑛 ⋅ 𝑟𝑒𝑐𝑎𝑙𝑙 = 𝑡𝑝 𝑝𝑟𝑒𝑐𝑖𝑠𝑖𝑜𝑛 + 𝑟𝑒𝑐𝑎𝑙𝑙 𝑡𝑝 + 0.5(𝑓𝑝 + 𝑓𝑛)

Мы будем учитывать баллы для всех отдельных классов, кроме тега O. Для объектов, состоящих из нескольких слов, мы используем “расслабленную” метрику: если границы прогнозируемого объекта совпадают с границами эталонного объекта, мы добавляем 1 к количеству TP (количество истинных положительных примеров). Если имеется только частичное совпадение, мы добавляем число от 0 до 1, вычисляемое как длина пересечения, разделенная на полную длину объекта.

Скрипты для оценки решения можно найти в папке evaluation репозитория semantic-role-labelling.

**Baseline решение**

Базовое решение находится врепозитории semantic-role-labelling.

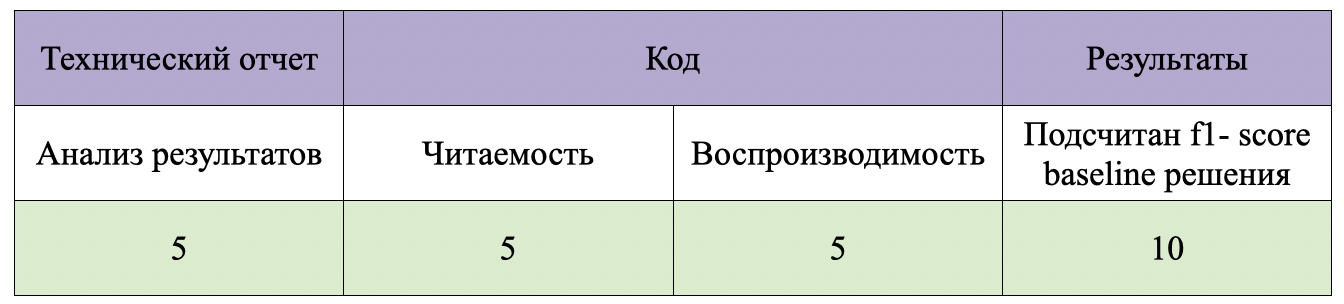
**Ожидаемый результат**

1. Воспроизведение baseline модели на Python 3 в виде

jupyter notebook (можно предоставить ссылку на Google Colab) с подсчетом f1-score с использованием скриптов для оценки решения (смотри **Метрики оценки качества**).

2. Описание полученного результата в том же jupyter notebook.

**Критерии оценки**

****

Чтобы получить 100% за задание, вам нужно набрать 25 баллов.