Решение задания 2

Основная задача разметки корпуса текстов состоит в том, чтобы сформировать обучающий материал, который может быть в дальнейшем либо для машинного обучения, либо для выведения эвристических правил. Поэтому чрезвычайно важно во-первых, четко определить список выделяемых объектов и их атрибутов, а во-вторых, описать то, как эти объекты и атрибуты должны выделяться разметчиком в тексте. В идеале инструкция разметчика должна позволять двум независимым экспертам порождать одинаково размеченный корпус.

А) Можно выделить четыре ключевых аспекта, из которых складывается оценка ресторана в целом:

1) Оценка еды – ее качества, размера порций, разнообразия меню. Это основная функция ресторана, то, что собственно составляет услугу, которую оценивают посетители

2) Оценка ресторана как помещения – впечатления посетителя от нахождения в пространстве ресторана: интерьер, атмосфера, чистота и пр.

3) Оценка качества обслуживания – коммуникация с официантами- то, что отличает ресторан от фастфуда и то, к чему посетители предъявляют особые требования

4)Цены: составив свое мнение о еде, интерьере и обслуживании в ресторане, посетитель соотносит свою оценку с тем, сколько он должен заплатить за полученные услуги, адекватность цен является еще одним оцень важным параметром оценки

Кроме того, добавим пятый аспект, в котором нет категориальной дифференциации – общее впечатление о ресторане.

Б) Разметка корпуса должна включать в себя, во-первых, выделение объектов, связанных с выделенными аспектами (Еда, Интерьер, Обслуживание, Цены, Общее), а также их атрибутов, содержащих собственно отрицательную, положительную или нейтральную оценку.

Ниже, в таблицу помещены лейблы объектов и значения их атрибутов

Лейбл объекта	Значение	Лейбл атрибута	Значение
MENU	Еда	MENU-POS/NEG/NEUTR	позитивное/негативное/нейтральное
INT	Интерьер	INT-POS/NEG/NEUTR	позитивное/негативное/нейтральное
SERV	Обслуживание	SERV-POS/NEG/NEUTR	позитивное/негативное/нейтральное
PRICE	Цены	PRICE-POS/NEG/NEUTR	позитивное/негативное/нейтральное
GEN	Общая оценка	GEN-POS/NEG/NEUTR	позитивное/негативное/нейтральное

Кроме того, введем отдельный атрибут anti, которые меняет полярность атрибута объекта на противоположную.

При разметке корпуса текстов следует соблюдать следующие правила:

1) Все лейблы и их атрибуты снабжаются сплошным индексом. Атрибут лейбла имеет тот же индекс, что и лейбл, например MENU1 и MENU-POS1
2) Формат записи индексов: сразу после лейбла или атрибута.
3) Объект и относящийся к нему атрибут выделяется с помощью квадратных скобок. Общий принцип состоит в том, чтобы помещать в скобки минимально возможный отрезок текста - с тем, чтобы минимизировать разнообразие используемых выражений и улучшить качество правил и машинного обучения. Выделяются существительные, непосредственно указывающие на объект, в любом падеже, без выделения предлогов в предложной группы. В качестве атрибутов выделяются качественные прилагательные без наречий, усиливающих их значение (например, очень) , или качественные наречия, или глаголы, или глагольная группа в том случае, если значение собственно глагола не несет в себе оценки
   Например:
   в [меню]MENU1 все [старое]MENU-NEG1

   Помимо того, что [официанты]SERV2 [забыли] SERV-NEG2 заказы двоих из гостей

   Сложные случаи и их разрешения
4) Если один кусок текста является атрибутом нескольких объектов, то значения атрибутов перечисляются через /
5) Если текст является и объектом, и атрибутом объекта атрибут (например, вкусно – это позитивная оценка меню, при этом сам объект выражен уже в этой оценке), то объект не выделяется, а выделяется сразу атрибут с индивидуальным индексом

Д)

Ниже перечислены основные проблемы, которые возникли в процессе разметки корпуса.

1) Разметка и структура объекта не предусматривает маркирования степени признака, хотя в корпусе имеются лексические выражения (например, очень) или графические (например, !!!!). Решение: было решено оставить структуру объектов без изменений, однако в будущем возможно предусмотреть градацию атрибутов по степени выраженности полярности
2) Значение некоторых атрибутов непонятно (например, «в стиле Новикова» в отзыве №2) Решение: было решено маркировать как нейтральное значение атрибута
3) В отзыве 11 встретился аспект местоложение, который не входит не в один из выделяемых объектов. Решение – новые аспекты добавлять к категории GEN
4) В отзыве 13 выстретилось два выражения, меняющих полярность, типа ANTI : к сожалению, на первый взгляд. Решение – каждое выражение было выделено отдельно, но снабжено общим индексом. Предполагается, что правило, ANTI работает только 1 раз в 1 клаузе.
5) Целый ряд предложений остался не разобранным. Есть три класса случаев.
   1-й случай:
   когда предложение не содержит в себе оценки , см. например отзыв 10:
   25 апреля заказала стол,через ресторан ру
   2-й случай:
   когда объект не относится к рассматриваемым аспектам и не может быть присоединен к категории общее, см. например, отзыв 5

Ваша вывеска режет глаз своей неграмотностью.

3-й случай:

оценка есть, но она не выражена эксплицитно, ср. отзыв 7

мои букеты роз (один из которых стоил 100 долларов и был куплен за космические деньги в Цветном) подарили другой имениннице вместе со всеми поздравительными открытками, которые я даже не успела прочитать

Поскольку исходная инструкция не задавала принципов разметки описанных случаев, эти предложения остались неразмеченными.