Манифест по разметке упоминаний объектов в комментариях

1. Определения

Объект мониторинга – конкретный объект реального мира, для которого может производиться мониторинг высказываний.

Упоминание объекта – часть сообщения, содержащая явное название объекта (в том числе сокращенные), либо слова/словосочетания, связанные кореферентной связью с названием объекта, то есть выражения (в том числе сленговые), подразумевающие конкретный объект, местоимения, заменяющие ранее упомянутый объект.

Явное упоминание объектов – последовательность слов, составляющих именную фразу (или прилагательное в значении существительного), значение (meaning) которой однозначно определено и соответствует значению, которое подразумевал автор текста. Под термин попадают:

- Термины
    - имена объектов (в том числе сокращенные), например, «Александр Турчинов» или «США»;
    - сленговые выражения, подразумевающие конкретный объект, например, «рашка»;
- местоимения, либо иная последовательность слов, связанная кореферентной связью с упоминанием объекта, например «собеседник» в интервью с конкретным человеком.

2. Разметка упоминаний объектов

В этом разделе будут рассмотрены вопросы, касающиеся определения границ и значений упоминаний объектов.

2.1. Границы явных упоминаний объектов

Границы явных упоминаний объектов рекомендуется выбирать, опираясь на следующие правила:

- В упоминание объекта попадают все не оценочные слова, описывающие объект;
- В упоминание объекта попадают слова, исключение которых меняет значение термина;
- В упоминание объекта должны попадать слова, приближающие термин к каноническому названию объекта. Например, словосочетание «Соединенные Штаты Америки» является термином, несмотря на то, что США часто называют «Соединенные Штаты» и смысл термина от этого не меняется;
- В упоминание объекта не должны попадать слова уточняющие значение упомянутого объекта, например, указание должности или звания описываемого человека.

Например,
в сообщении «МЧС по Боровску Калужской области заявило …» необходимо выделить явное упоминание «МЧС по Боровску Калужской области».

2.2. Анафоры

Под упоминания объектов также попадают местоимения и другие слова, связанные с прошлым упоминанием объекта в виде анафорического отношения [https://ru.wikipedia.org/wiki/Анафора_(лингвистика)]. Такие упоминания объектов следует размечать по стандартным правилам, указывая специальным флажком (см. Рисунок 1), что упоминание не является самостоятельным, а связано с каким-то другим упоминанием объекта.

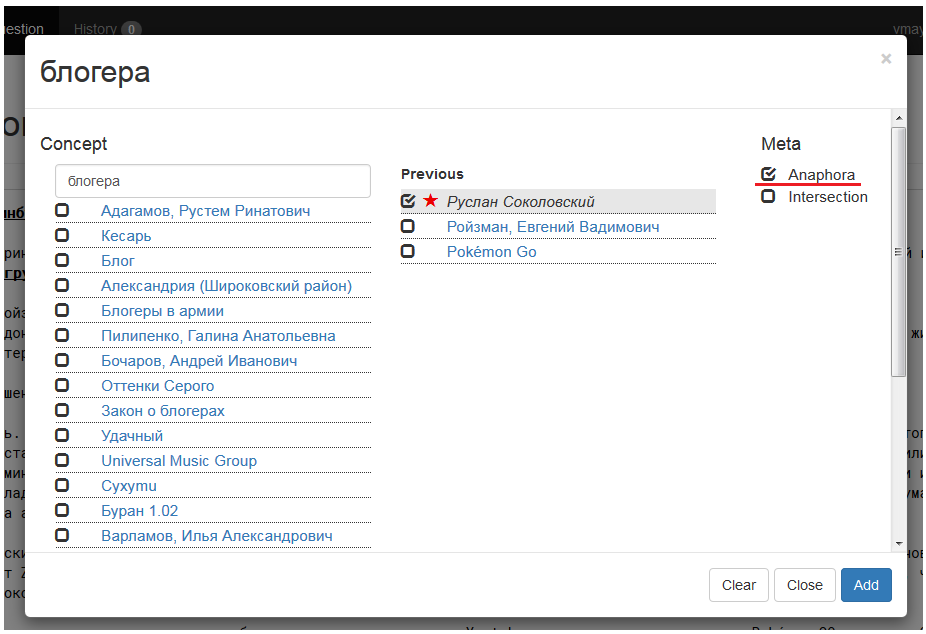

Рисунок 1. Выделение анафорического выражения

2.3. Значение упоминания объекта

Значения объектов бывают двух типов: **точное значение** упоминания объекта (значение – непосредственно упомянутый объект) и **косвенное** (значение - сильно связанный ассоциативно с точным значением).

Прежде всего, необходимо выделять точные значения (как правило, одно) явно помечая их точными значениями (звездочка слева от названия объекта) (см. Рисунок 2). Для ускорения разметки первое выделенное значение автоматически помечается точным.

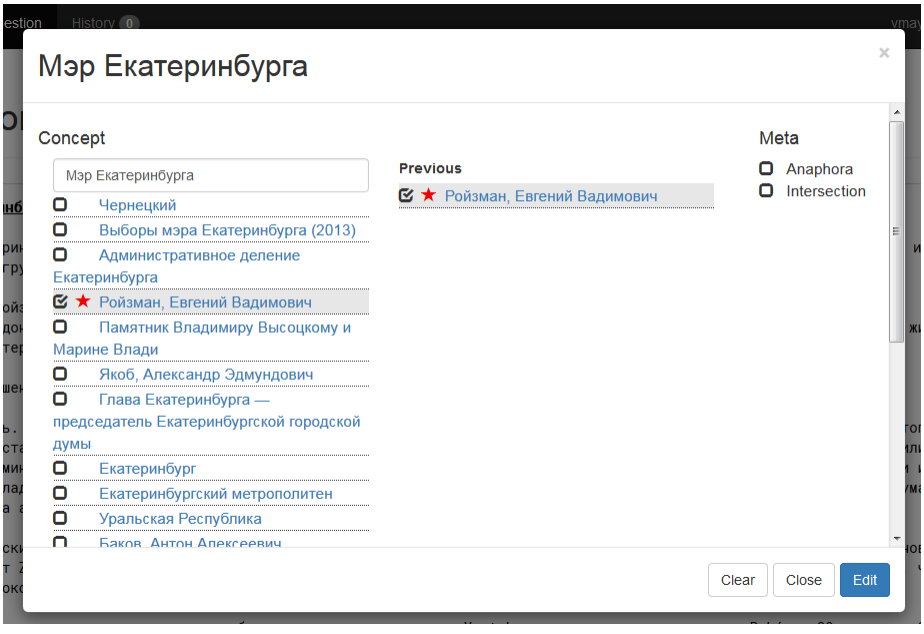

Рисунок 2 Выделение точного значения

Как правило, выбор конкретного значения (особенно это касается упоминаний людей) не вызывает особой сложности.

При разметке точных значений упоминаний объектов необходимо выделять именно то значение, которое подразумевается в контексте фразы или диалога.

Например,
явному упоминанию «Президент РФ» должно быть поставлено в соответствие значение «Путин, Владимир Владимирович», если речь идет про конкретного человека (часто). Значение «Президент Российской Федерации» должно быть отмечено как косвенные значения упоминания объекта.

При разметке значений упоминаний объектов необходимо учитывать и время, про которое идет речь в сообщении. Например, явному упоминанию «Президент РФ» должно быть поставлено в соответствие значение «Медведев, Дмитрий Анатольевич», в случае если в сообщении речь идет про 2011 год.

2.4. Упоминание нескольких объектов

Довольно часто бывают ситуации, когда упоминанию объекта соответствует более одного точного значения. В таком случае необходимо размечать все точные значения, соответствующие упоминанию.

Например,
упоминанию «главы государств» в обсуждении Российско-Американских отношений следует сопоставить точные значения «Путин, Владимир Владимирович» и «Обама, Барак».

В случае если точных значений необходимо разметить слишком много (более 5), например, «депутаты ГД от Единой России», следует выбирать наиболее близкое значение, охватывающее все предполагаемые значения (в приведенном примере – одно значение «Единая Россия»).

2.5. Пересекающиеся упоминания объектов

Если при разметке встретились два или более пересекающихся упоминаний объектов, необходимо выделить одно упоминание объекта (объединенное) и назначить ему несколько значений, соответствующих каждому из упоминаний, входящему в объединенное упоминание объекта. Для объединенного упоминания объекта необходимо явно указать свойство «intersection».

Например,
весь фрагмент сообщения «Донецкая и Луганская народные республики» необходимо выделить в качестве одного объединенного термина и поставить в соответствие ему значения «Донецкая народная республика» и «Луганская народная республика».

2.6. Отсутствие значения упоминания объекта в базе знаний

Для большого количества упоминаний объектов в используемой базе знаний не содержится действительного значения объекта (ситуация особенно распространена для событий). Такие упоминания должны помечаться специальной константой **NOT_IN_KB**. Дополнительно рекомендуется указывать каноническое имя объекта (в стиле Википедии).

Обратите внимание, что константа **NOT_IN_KB** должна указываться только для точных значений упоминаний объектов, в случае отсутствия косвенного значения в базе знаний, это значение необходимо пропускать.

В случае если упоминанию должны соответствовать несколько значений, остальные значения размечаются по стандартным правилам.

2.7. Метонимия

Особое внимание стоит обратить на те упоминания объектов, значения которых не соответствуют общепринятым. Например, «Москва заявила …»: «Москва» является упоминанием не города, а какого-то человека или органа государственной власти (зависит от контекста).

Подобные упоминания объектов необходимо размечать, руководствуясь следующими соглашениями (сформулированы для наиболее частых упоминаний).

2.7.1. Упоминание государств

- Упоминание, выраженное названием государства («Россия», «США») → высший исполнительный орган соответствующего государства («Правительство РФ», «Президент США»).

2.7.2. Упоминание городов

- упоминание, выраженное названием столицы государства («Москва», «Вашингтон»), когда речь идет про государство в целом → эквивалентно упоминанию государства (см. предыдущий пункт);
- упоминание, выраженное названием столицы региона («Москва»), когда речь идет о регионе в целом → главный орган исполнительной власти региона («Правительство Москвы»);
- упоминание города, в котором расположена межгосударственная организация («Брюссель») → соответствующая межгосударственная организация («Европейский союз»).

2.7.3. Упоминание именованных зданий/географических мест

- Упоминание, выраженное названием здания, в котором находится организация («Кремль», «Пентагон») → организация (наиболее известная), расположенная в соответствующем здании («Администрация президента России», «Министерство обороны США»);
- Упоминание, выраженное полным, или частичным адресом («Лубянка», «Петровка 38») → организация, расположенная по этому адресу («Федеральная служба безопасности», «Главное управление МВД России по городу Москве»).