Первое информационное письмо FactRuEval

Уважаемые коллеги! В этом году в рамках Dialogue Evaluation будет проводиться соревнование по извлечению информации из новостных текстов на русском языке. Участникам будут предложены дорожки по **извлечению именованных сущностей** и по **извлечению фактов**.

В 2004-2006 гг. в рамках семинара РОМИП (www.romip.ru) уже проводились дорожки по фактографическому поиску, где извлечение именованных сущностей и фактов было в центре внимания. С тех пор прошло более 10 лет. За это время специалистами в области компьютерной лингвистики, интеллектуального анализа данных и других смежных направлений разработаны новые методы, которые реализуются для практически значимых приложений, как большими компаниями, так и небольшими исследовательскими группами. Вместе с тем, в российской действительности мало достоверной информации о результатах, полученных и теми, и другими в данной области. Вот почему независимое сравнительное тестирование систем извлечения информации для русского языка было бы не только интересно всем исследователям и разработчикам, но и крайне полезно для дальнейшего развития методов и систем обработки естественного языка.

Приглашаем всех желающих принять участие в независимом тестировании систем извлечения информации. В качестве участников мы будем рады видеть как опытных разработчиков, так и новичков, в том числе и тех, кто раньше не занимался этой областью.

В рамках соревнования будет предложено три дорожки, для каждой из которых будет предоставлено описание стандарта разметки, а также обучающая коллекция размеченных текстов. Коллекция будет состоять из новостных текстов на русском языке, опубликованных в Интернете в 2005-2015 годах.

1. **Дорожка по извлечению именованных сущностей**

Это классическая задача NER: требуется выделить в текстах именованные сущности определенных типов (персоны, организации и локации).

2. **Дорожка по извлечению именованных сущностей с выделением атрибутов**

Это усложненный вариант предыдущей дорожки, в которой, помимо определения границ упоминания сущности в исходном тексте, требуется выделять строковые атрибуты этой сущности в нормализованном виде. Для персон это фамилия, имя и отчество. Для организаций и локаций – нормализованное название (одно или несколько).

3. **Дорожка по извлечению фактов**

В этой дорожке требуется выделить факты определенных типов. Под фактом в данном случае подразумевается тип факта (например: “встреча”, “покупка”, ...) и набор строковых полей (например: “участник встречи 1”, “участник встречи 2”, “место встречи”, “дата/время начала встречи”, …). Факт считается правильно выделенным, если у него правильно заполнены все поля (актанты), но частичное их выделение также будет учитываться. Существенное отличие от дорожек по выделению именованных сущностей заключается в том, что здесь не требуется привязка факта к позиции в тексте, т.е. неважно, из какой именно фразы был выделен тот или иной факт, а необходимо только обнаружить факт определённого типа и правильно заполнить его поля.

Планируется разметка коллекции фактами следующих типов: Occupation (персона, должность, организация), Meeting (список участников, место), Purchase&Sale (список участников сделки). Однако этот список может быть скорректирован с учетом интересов участников (напишите нам).

4. Дополнительные дорожки

Возможно, что в соревнование будут включены дополнительные дорожки и дополнительные коллекции текстов. Для того, чтобы это произошло, напишите нам письмо с описанием дорожки, в проведении которой вы заинтересованны как потенциальный участник или как потенциальный пользователь или заказчик. Мы будем рады сотрудничеству.

Подробные инструкции (соглашения) по выделению каждого из упомянутых типов сущностей и фактов на данный момент находятся в разработке.

Ближайшие шаги

В настоящий момент нам (оргкомитету) важно оценить количество потенциальных участников соревнования и их интересы. Если вы рассматриваете возможность принять участие в одной из заявленных дорожек, заполните Заявку потенциального участника. Если вы хотите предложить свою дорожку, заполните следующую Форму. Вы также можете написать письмо Гусевой Анне по адресу: evaluation@dialog-21.ru.

Мы будем рады не только заявкам, но и любым конструктивным комментариям и предложениям, связанным с проведением конкретных дорожек (принципы разметки, оценки качества и т.п.).

Предполагаемый график проведения соревнования

10.11.15 Объявление дорожек

10.12.15 Публикация обучающей коллекции

11.01.16 Публикация коллекции для оценки результатов

17.01.16 Прием результатов прогонов

29.01.16 Объявление результатов участникам

15.02.16 Подача статьи на конференцию Диалог

С уважением,

Оргкомитет FactRuEval

Над организацией мероприятия работает инициативная группа в составе:

Виктор Бочаров (OpenCorpora)

Ирина Ефименко (НИУ ВШЭ)

Анатолий Старостин (ABBYY)

Мария Степанова (ABBYY)

Светлана Толдова (НИУ ВШЭ)

Владимир Хорошевский (ВЦ РАН)

Анна Гусева (Диалог)