Второе информационное письмо. MorphoRuEval-2017, Dialogue Evaluation

Соревнование по оценке методов морфологического анализа русского языка

В 2016-2017 гг. планируется организовать соревнование по оценке методов морфологического анализа русского языка MorphoRuEval.

Цель соревнования — стимулировать развитие технологий морфологической обработки текстов на русском языке, в особенности текстов из сети Интернет, как нормативных (новости, литературные тексты), так и менее формального характера (блоги и другие социальные медиа). Задача корректной обработки таких текстов является одной из наиболее приоритетных задач в современной прикладной лингвистике, так как объем неформальных текстов все время растет за счет социальных сетей, а вся дальнейшая цепочка обработки текста сильно зависит от результатов морфологического анализа. Особенно важной эта проблема становится при сборе корпусов из сети Интернет и последующих лингвистических исследованиях на их материале.

Командам предлагаются следующие дорожки:

1. с закрытым набором данных, участникам которой разрешено обучать свои модели только на предоставленных данных (это интересно, прежде всего, научно-исследовательским группам и студенческим командам, у которым нет своих больших корпусов)

Для верификации результатов участники закрытой дорожки должны разместить свой код на github, чтобы он был публично доступен - это удобно как организаторам, так и другим командам-участникам.

2. с открытым набором данных, участникам которой разрешается использовать любые данные, в том числе данные соревнования.

Обучение систем:

Подготовка данных занимает большую часть ресурсов оргкомитета, и, чтобы достигнуть компромисса между высоким качеством размеченных данных и удобством участников дорожки, предлагается работа с данными по следующему сценарию: 20-21 января всем участникам рассылаются данные в предварительном формате, однако с некоторыми описанными особенностями и различиями. В течение нескольких недель (ориентировочно до 5 февраля) оргкомитет принимает замечания и пожелания участников по поводу формата, тестирует и верифицирует разметку, и затем в середине предоставляет тестовую выборку в окончательном задокументированном формате. К данным прилагается подробное описание разметки: списки закрытых классов слов для русского языка, а также списки слов, разметка которых представляется оргкомитету настолько спорной, что не будет учитываться при подсчете итоговых метрик.

Материал, предоставляемый участникам соревнования для обучения:

Неразмеченные корпуса (длина в словах):

- Живой Журнал ГИКРЯ, 30 млн
- Современная художественная проза, 300 млн
- Социальные сети, 10 млн (ВКонтакте, Фейсбук)

Размеченные данные, в морфологической кодировке UD 1.4 и 2.0:

- материал OpenCorpora
- Живой Журнал ГИКРЯ
- Корпус со снятой омонимией НКРЯ

О размере размеченной обучающей и контрольной выборки будет сообщено дополнительно, т. к. мы хотим проводить соревнование только на данных с верифицированной разметки.

Размеченные материалы унифицированы относительно критериев оценки (см. раздел «Процедура оценивания»), но могут различаться в некоторых решениях относительно разрядов союзов и т.д., не влияющих на оценку. Материалы предоставляются в формате UD 1.4 и 2.0. Различия форматов для русского языка:

1. VerbForm=Trans → VerbForm=Conv
2. CONJ→CCONJ
3. Voice=Mid для глаголов на -ся/-сь

Полный список изменений: http://universaldependencies.org/v2/summary.html

Официальный морфологический стандарт разметки – Universal Dependencies 2.0, и мы просим участников предоставлять размеченные коллекции только в этом формате (в том же формате, в котором будет выдана обучающая выборка).

В качестве контрольной выборки участникам будет предоставлен токенизированный материал без разметки. Токенизацию необходимо сохранить при сдаче решения организаторам.

Обновленный график проведения соревнования:

- 20-25 января – второе информационное письмо, пример размеченных данных, открытие регистрации
- 10-15 февраля – обучающая выборка с учётом замечаний участников по формату
- 1 марта – контрольная выборка, рассылка (5 дней для участников)
- 15 марта – публикация результатов, дедлайн подачи финальной версии статей

Подача статей участников соревнования:

Участники по желанию могут подавать статьи о своих разработках на конференцию «Диалог» в рамках соревнования, в таком случае участники должны заранее заявить о своем намерении подать статью оргкомитету, чтобы получить отсрочку. Возможна подача статьи на рецензирование в 2 этапа – в феврале основной текст статьи, без

итогов соревнования, а в начале марта – исправленный вариант с итогами дорожки. Однако дедлайн подачи объявленных статей – начало марта.

Процедура оценивания:

В качестве контрольной выборки участники получат токенизированный материал, который будет необходимо разметить и лемматизировать согласно формату UD 2.0 или 1.4.

Оценке подлежат граммемы категорий, указанные в столбце «оцениваемые», при этом оценка не будет различать положительную и превосходную степени сравнения у прилагательных, причастия должны считаться прилагательными с соответствующей леммой (проявлявших → проявлявший), а большинство предикативов считается наречиями. О метриках качества будет сообщено дополнительно.

Помимо качества определения метки части речи оценивается:
1) Существительное: род, число, падеж
2) Прилагательное: род, число, падеж, краткость, сравнительная степень
3) Глагол: наклонение, лицо, время, число, род.
4) Местоимение: род, число, падеж, лицо, синтаксический тип (в разметке UD 2.0 большинство местоимений, играющих роль предмета, относятся к категории PRON, а играющих роль определения – к категории DET).
5) Наречие: сравнительная степень
6) Числительное: род, число, падеж, графическая форма (при этом порядковые числительные считаются прилагательными).

Не оценивается частеречная разметка предлогов, союзов, частиц, междометий и "остального". Для этих частей речи предоставляются конечные списки. Также предоставляется закрытый список местоимений (категории PRON и DET).

Также в обучающей выборке размечена одушевлённость существительных и залог глаголов, который при этом является чисто графической категории, различающей глаголы с возвратным постфиксом -ся/-сь и остальные.

ОЦЕНИВАЕМЫЕ			
часть речи	**оцениваемые**	**размечаемые**	**комментарий**
существительное	род		
	число		
	падеж		
прилагательное	род		Причастия оцениваются вместе с прилагательными
	число		
	падеж		
	краткость		
	сравнительная степень		не разделяется превосходная и положительная

глагол	наклонение	залог	
	лицо		
	время		
	число		
местоимение	род	класс	
	число		
	падеж		
	синт. тип		
	лицо		
наречие	сравнительная степень		большинство предикативов считается наречиями
числительное	род		
	число		
	падеж		
	графический тип		
РАЗМЕЧАЕМЫЕ			
предлог			
союз			
частица			
междометие			
остальное			

Репозиторий соревнования: https://github.com/dialogue-evaluation/morphoRuEval-2017/

Отображение между наборами категорий для русского языка, добавлен формат UD: https://github.com/kmike/russian-tagsets

Организаторы:

Алексей Сорокин (МГУ, МФТИ, ГИКРЯ), Татьяна Шаврина (ГИКРЯ, НИУ ВШЭ), Алексей Зобнин (Яндекс), Ольга Ляшевская (НИУ ВШЭ), Виктор Бочаров (Яндекс, Открытый корпус), Светлана Алексеева (Открытый корпус), Кира Дроганова (Charles University), Алена Феногенова (НИУ ВШЭ, НИИ КВАНТ), Илья Карпов (НИИ КВАНТ, НИУ ВШЭ), Даниил Селегей (ABBYY, ГИКРЯ, РГГУ).

Контактный адрес: geekrya@gmail.com

С уважением,

организаторы соревнования