ОЦЕНКА МЕТОДОВ АВТОМАТИЧЕСКОГО АНАЛИЗА ТЕКСТА: МОРФОЛОГИЧЕСКИЕ ПАРСЕРЫ РУССКОГО ЯЗЫКА

NLP EVALUATION: RUSSIAN MORPHOLOGICAL PARSERS

Ляшевская О., *Университет Тромсё, Норвегия,* ***Бонч-Осмоловская А., Толдова С.,*** ***Астафьева И., Гарейшина А., Гришина Ю., Дьячков В., Ионов М., Королева А., Кудринский М., Литягина А., Лучина Е., Сидорова Е.,*** *МГУ им. М.В.Ломоносова,* ***Савчук С.,*** *ИРЯ РАН им. В.В.Виноградова,* ***Коваль С.*** *(lingtecheval@yahoo.com)*

Форум «Оценка методов АОТ» (http://ru-eval.ru) – новая инициатива, целью которой является независимая оценка методов и алгоритмов работы русскоязычных лингвистических ресурсов. В статье описываются принципы и процедура проведения дорожек форума, состав участников, тестовая коллекция, организация экспертизы и полученные результаты.

1. Введение

Форум «Оценка методов автоматического анализа текста» стартовал в феврале 2010 года, и темой первого года стали морфологические парсеры русского языка. Тестовый запуск систем и экспертиза ответов состоялись в марте–апреле, а очную встречу участников и обсуждение результатов предполагается провести на конференции «Диалог'2010». Сама идея форума возникла два года назад на конференции Language Resources and Evaluation (LREC'2008), но настоящим своим рождением она обязана конференции «Диалог». Вместе с постоянными участниками «Диалога» мы обсуждали, отчего, несмотря на существование старых и хорошо зарекомендовавших себя парсеров русского языка, все время появляются новые процессоры, нужны ли лингвистические ресурсы, например, словари, для построения компьютерных лингвистических систем, в чем задача лингвистов на разных этапах развития IT–технологий и, наконец, почему в мире большой популярностью пользуются некоммерческие семинары по сравнительной оценке парсеров (ср. проекты CLEF, AMALGAM, GRACE, EVALITA, SEMEVAL и др.) и не нужно ли ввести такую моду в России для русскоязычных ресурсов.

Ключевое событие форума строится в игровой форме – системы соревнуются друг с другом на специально подготовленной коллекции текстов, кто даст больше правильных ответов. Однако цель соревнования вовсе не в том, чтобы назвать победителя, а в том, чтобы выявить, какие алгоритмы и ресурсы позволяют улучшить результаты по тому или иному показателю. В связи с этим форум предполагается проводить регулярно, чтобы дать разработчикам возможность из года в год совершенствовать свои методы. Таким образом, настоящая высокая цель форума – улучшение состояния науки в области автоматической обработки текста. Но главное, форум должен способствовать созданию среды, в которой научные, научно–производственные, коммерческие разработки могли бы проходить независимую экспертизу, и в которой могли бы обсуждаться проблемы и перспективы развития технологий.

Немаловажным представляется и практический выход, полученный по окончании данного соревнования: корпус вручную размеченных и выверенных текстов, который можно использовать в научно–исследовательских целях, сформированные принципы разметки, к которой могут быть приведены разметки большинства систем, исчисление сложных случаев русского языка, которые не имеют однозначного решения. Счастливым образом, форум 2010 года получил также образовательную составляющую – в его подготовке, проведении и формировании финального отчета активное участие принимали студенты Отделения теоретической и прикладной лингвистики филологического факультета МГУ

им. М.В.Ломоносова, которые получили возможность «пощупать руками», как работают парсеры, увидеть, в чем их сильные и слабые стороны, чем парсеры системно отличаются друг от друга и т.д.

Объектом рассмотрения в данном форуме являются не собственно морфоанализаторы, работающие с изолированными словами (именно они рассматривались в качестве объекта оценки в отдельных работах последнего десятилетия, ср. Коваль 2003), а модули, учитывающие или потенциально учитывающие контекст. В связи с этим как в названии форума, так и во всей его внутренней документации последовательно используется понятие «морфологический парсер», обозначающее модуль, функциональность которого позволяет, как минимум, обрабатывать сразу всю текстовую цепь слов, и как максимум, учитывать при анализе каждого текстового слова результаты разбора его соседей. В этой второй, «сильной» интерпретации термин «морфологический парсер» становится практически неотличимым от используемого в англоязычной литературе «POS tagger», однако организаторы форума предпочитают говорить о «морфологических парсерах» в силу специфики русского языка: как «слабые» (не предусматривающие контекстную дизамбигуацию разборов), так и «сильные» (включающие такую дизамбигуацию) варианты парсеров опираются на заложенную их разработчиками модель такого далеко не тривиального объекта, как русская словоизменительная морфология, а, значит, имеют достаточно много общего.

Важнейшая презумпция организации соревнования состояла в том, что не бывает единственно правильного решения грамматически спорных вопросов и единственно правильного алгоритма морфологического анализа. Существует множество примеров того, как оптимальный выбор того или иного решения зависит от той цели, для которой проводится анализ. Так, выделение устойчивых словосочетаний как одной единицы (например, «Государственная Дума») может улучшить качество информационного поиска, двукомпонентный анализ, в данном случае, необходим для корректных последующих уровней обработки. Разбор словоформы «бело–кремовое» как единого целого, получающего грамматическую характеристику по концовке, вполне удовлетворителен во многих ситуациях, однако для тех систем, в цикл обработки которых включен семантический анализ, для осмысления этой явно несловарной формы наверняка потребуется ее сегментация по дефису.

В связи с этим достаточно широкий круг грамматических вопросов был вынесен за скобки соревнования и не оценивался. Тем не менее, именно эти проблемы – и расхождения систем в предлагаемых решениях – явились предметом особого внимания со стороны организаторов. Нам представляется, что исчисление и классификация случаев, сложных для автоматического грамматического разбора, а также сведения о частотности возможных решений являются самоценной информацией, которая может быть использована научным сообществом и для исследовательских целей, и для улучшения эффективности прикладных разработок.

2. Дорожки

Организационно форум 2010 года во многом строился по образцу Семинара по оценке методов информационного поиска РОМИП (РОМИП 2009). Оценка алгоритмов проводилась по нескольким независимым дисциплинам (дорожкам). Каждая дорожка была посвящена одной конкретной задаче анализа текста с заранее согласованными правилами оценки систем–участников. От участников не требовалось участия во всех дорожках сразу, поэтому у них была возможность сосредоточиться на решении только одной из предлагаемых задач.

В соревнованиях рассматривались два типа морфологических разборов:
1) без дизамбигуации: системы дают множество возможных разборов, оценивается наличие среди них правильного разбора;
2) с дизамбигуацией: система должна дать единственный правильный разбор, корректность которого является объектом оценки.

Соревнования без дизамбигуации состоялись на следующих дорожках:
- «Лемматизация». Задача этой дорожки состояла в том, чтобы правильно определить исходную форму словоформы
- «POS». Требовалось правильно определить часть речи, к которой принадлежит исходная словоформа
- «Морфология». Задача: правильно определить грамматические теги, которые характеризуют исходную словоформу, например, род, число, падеж, время и т.д. Оценивалось наличие правильной комбинации грамматических тегов, представленных в разборе
- «Редкие слова». Задача состояла в том, чтобы правильно определить лемму и часть речи для списка специально отобранных несловарных или нестандартных словоформ.

Соревнования с дизамбигуацией проводились на дорожках
- «Дизамбигуация: леммы» и
- «Дизамбигуация: POS».[1]

3. Участники

На конкурс были поданы заявки от 15 групп разработчиков из Москвы, Санкт-Петербурга, Екатеринбурга (Россия), Минска (Беларусь), Донецка (Украина), Лидса (Великобритания). В тестовых дорожках приняли участие 12 систем: ARME, Crosslator, FSTMorph (+ ЭТАП-3), Libmorphrus, Mocky, Mystem (+ FastDictionary), Polymorph, Pymorphy, RDMA_IAI, Semantarus Morpho, Starling, TextAn[2]. Некоторые разработчики представляли несколько вариантов морфологических анализаторов для дорожек с дизамбигуацией и без нее и даже несколько вариантов реализации алгоритмов на одной дорожке.

В итоге было получено 13 ответов систем по дорожкам «Лемматизация» и «POS», 12 ответов по дорожке «Морфология», 8 ответов по дорожке «Редкие слова» и 7 ответов по обеим дорожкам с дизамбигуацией. Ответы одного участника по дорожкам «Лемматизация», «POS» и «Морфология» были дисквалифицированы за несоответствие формата данных и не участвовали в экспертизе.

4. Тестовая коллекция и задания

Для соревнования была подготовлена общая коллекция неразмеченных текстов для дорожек «Лемматизация», «POS», «Морфология», «Дизамбигуация: леммы» и «Дизамбигуация: POS» (Основная коллекция) объемом около 1 млн. словоупотреблений. Материалы для Основной коллекции были составлены из фрагментов текстов, присланных некоторыми участниками и экспертами. В Основную коллекцию вошли тексты различной тематики и жанровой принадлежности в следующих соотношениях:

---

[1] Первоначально предполагалось также проведение дорожки «Коллекции: Грязные тексты», где системам ставилась задача разметить фрагменты плохо распознанных отсканированных документов, таблиц, содержащих слова с некорректно внесенными знаками переносов и форматирования и текстов с большим количеством опечаток. Была подготовлена и разослана участникам специальная коллекция, однако, поскольку по этой дорожке был получен только один ответ, дорожка была отменена и экспертиза результатов по ней не проводилась.

[2] Еще одна система (АОТ) выступала вне зачета, с согласия автора ее запускали студенты-эксперты. Более подробную информацию об участниках можно найти на странице http://ru-eval.ru/participants.html.

18% Статьи в СМИ/Нон-фикшн, 15% Новости; 15% Интервью; 15% Технические тексты; 15% Юридические тексты; 18% Художественная литература; 4% Блоги и форумы.

На базе Основной коллекции было составлено задание для дорожки «Редкие слова», включавшее 75 отобранных экспертами слов с их ближайшим контекстом, в том числе:
1) продуктивные модели (слова с неизвестным словарю корнем, но образованные с помощью продуктивных аффиксов. Среди них встречаются так называемые слова–обманки: *аррабьята* (лемма «аррабьята») vs *френдята* (лемма «френденок») и т. п., а также авторские «придуманные» слова: *увазила, кругтелся, склипких, грезитвой*;
2) сложные слова, у которых вторая часть совпадает со словами или вторыми частями сложных слов в словаре Зализняка: *полуколебаний*, *ультраженственной, миллионометра, Росторгмонтаж*;
3) слова с «неизвестными» корнями (в т. ч. имена собственные), не содержащие продуктивных аффиксов, для которых носители языка могут однозначно определить лемму и часть речи (по стандартным окончаниям русского языка и зная контексты, в которых они употреблены): *турбийона* (лемма «турбийон»), *френдя* («френдить»), *тюрбо* («тюрбо»), *Баухаус* («Баухаус») и др.;
4) редкие и нестандартные формы (некоторые деепричастия, формы первого лица глаголов и степени сравнения, которые употребляются в языке, но признаются окказиональными или ненормативными, в связи с чем обычно отсутствуют в словарях): *стригя, пья, побежу*, *висю*, *деревяннее, нельзей*;
5) Аббревиатуры типа *ВЧК*, *ОГПУ, МФТИ*, которые система могла бы спутать с глаголами или словами других классов и ошибиться в определении леммы.

Источником выборки редких слов послужили научные тексты, инструкции, кулинарные рецепты и меню, записи речи детей дошкольного возраста (большинство интересных продуктивных моделей и нестандартных форм было обнаружено именно там, поскольку в возрасте с 3 до 5 лет дети постоянно изобретают новые слова), форумы в Интернете, а также тексты Велимира Хлебникова и Людмилы Петрушевской. Итоговый баланс задания «Редкие слова» включает 27 существительных, 13 прилагательных, 28 глаголов и 7 слов категории ADV.

Сравнение результатов по всем дорожкам проводилось на основе выборочной проверки ответов систем–участников. Для этого был подготовлен «Золотой Стандарт» – множество случайно выбранных предложений из Основной коллекции, объемом около 2000 словоупотреблений. В ходе экспертизы ответы систем сравнивались с произведенной экспертами ручной разметкой Золотого Стандарта, см. п. 6.

5. Принятые соглашения по унификации грамматической информации

Подготовительный этап потребовал определенных решений, направленных на унификацию нотации и структуры морфологических разборов в ответах, ожидаемых от парсеров. Было выявлено несколько типов проблемных случаев:
1) некоторые частеречные категории не имеют устойчивой общепринятой нотации разметки и выделяются, обозначаются и объединяются системами по–разному, что может затруднить оценку результатов (например, в одних системах выделяется один общий класс местоимений, в других системах они разводятся по классам существительных, прилагательных, наречий и т.д., в третьем случае выделяются классы местоимений–существительных, местоимений–прилагательных и т.п.);
2) объем парадигмы может разниться от системы к системе, например, формы парных глаголов совершенного и несовершенного вида могут приводиться к двум разным леммам (*прыгнул – прыгнуть*, *прыгал – прыгать*) или к одной общей (*прыгать*); часто

само требование к объему парадигмы зависит от того, для решения какой прикладной задачи используется модуль морфологического парсинга;
3) некоторые классифицирующие признаки словоформ (например, переходность у глаголов) могут считаться избыточными на этапе морфологического анализа текста, а их определение может быть затруднено в том случае, если анализируемая словоформа не входит в словарь системы;
4) некоторые морфологические признаки не могут быть однозначно определены в рамках морфологического анализа (например, нетривиально определение леммы и залога для глаголов с постфиксом *–ся*);
5) некоторые морфологические характеристики (например, звательный падеж) имеются только у ограниченного числа словоформ и могут системно не выделяться.

С учетом ожидаемых расхождений было принято решение о том, что разметка будет производиться парсерами по упрощенной системе. При лемматизации буквы *е* и *ё*, а также написание с прописной/строчной буквы признавались равноправными. Частеречные признаки были приведены к следующему сокращенному инвентарю: существительные (S), прилагательные (A), глаголы, в том числе причастия и деепричастия (V), предлоги (PR), союзы (CONJ), и сборная категория, включающая прочие несклоняемые слова: наречия, вводные слова, частицы, междометия (ADV). Не участвовали в оценке и могли быть размечены любым образом местоимения (включая наречные и предикативные), числительные, а также составные предлоги и союзы (ср. *потому что, в течение*).

Кроме того, был сокращен и список грамматических характеристик, приписываемых словоформе. В общем случае, сопутствующий набор грамматических признаков определялся тем минимумом информации, который нужно знать для однозначного восстановления словоформы из леммы. Морфологические признаки указывались только для существительных, глаголов и прилагательных.

Итоговый список размечаемых морфологических характеристик словоформ включает:

- род: m (мужской), f (женский), n – (средний)
- падеж: nom (именительный), gen (родительный, в том числе счетная форма – два шар/а), dat (дательный), acc (винительный), ins (творительный), loc (предложный, в том числе второй предложный, ср. *в лесу*)
- число: sg (единственное), pl (множественное)
- время: pres (= непрошедшее: настоящее и будущее время – *пишу, напишу*), past (прошедшее),
- наклонение: imper (повелительное)
- инфинитив: inf
- причастие: partcp,
- деепричастие: ger
- залог: act (действительный), pass (страдательный) – указывается только в формах причастий
- лицо: 1p, 2p, 3p

Таким образом, из классифицирующих категорий необходимым для указания являлся только род, не рассматривались переходность и вид глагола, залог для всех форм глагола кроме причастий и деепричастий, одушевленность имен. Кроме того, необязательно было указывать при разборе степень сравнения прилагательных и наречий, а также полноту/краткость прилагательных.

Следует также отметить, что не участвовал в оценке целый ряд непродуктивных словоизменительных категорий, а также маргинальных реализаций продуктивных категорий: лицо и наклонение форм императива 1 лица типа *пойдемте*; падеж имен в

конструкциях «пойти в *солдаты*», «попить *чаю*»; звательный падеж (*Маш! отче* и др.); род слов общего рода (*врач*).

6. Подготовка Золотого Стандарта

Ручная разметка Золотого Стандарта, предшествовавшая экспертизе результатов, преследовала несколько целей. Во–первых, требовалось независимое основание для автоматического сопоставления ответов систем, которое уменьшило бы объем ручной экспертизы: проверке подлежали только случаи расхождения между стандартом и ответами систем. Во–вторых, организаторы хотели избежать влияния результатов, предоставленных системой, на интуицию экспертов, и пропусков ошибок по невнимательности. В–третьих, разметка Стандарта должна была подготовить экспертов к оценке ответов систем, сформировать у них представление о том, какие сложные случаи их ожидают, понять объективную природу несовпадения некоторых ответов и выработать критерии для их либеральной оценки.

В разметке Стандарта принимало участие 10 экспертов, каждый фрагмент размечался независимо двумя разметчиками. Перед ними стояла задача выделить в тексте все русские словоформы и дать им единственный разбор. После технической валидации разметки на предмет соблюдения формата и допустимых сочетаний тегов согласованность результатов ручной разметки (inter–annotator agreement) составила: леммы – 94.4%, POS – 95.4%, морфология – 89.0%, весь разбор в целом – 85.5%. Оставшиеся содержательные расхождения согласовывались экспертами в паре. В случае если эксперты не могли прийти к единому решению, спорные вопросы выносились на обсуждение на специально организованных семинарах с участием всех разметчиков и еще 5 экспертов. В частности, обсуждалось, как лемматизировать потенциальные pluralia tantum, сокращения, слова с дефисом или незнакомые слова; к какому классу принадлежат слова типа *минувший*: причастие или отпричастное прилагательное; *данные*: прилагательное или отадъективное существительное? Каждый эксперт высказывал свое мнение по поводу того или иного случая, а также объяснял свою точку зрения. Затем наиболее убедительное решение вносилось в Золотой Стандарт. Например, в случае выбора леммы для *72–часовых* было предложено три возможных решения: 1) это две словоформы, которым приписываются две леммы: «72» и «часовой»; 2) лемма – «72–часовой»; 3) лемма – «семидесятидвухчасовой». В ходе дискуссии предпочтение было отдано первому варианту, который и был отражен в Золотом Стандарте.

7. Экспертиза ответов систем

Процедура экспертизы ответов морфологических парсеров предусматривала сравнение разбора каждой входящей в зачет словоформы с ее разбором в Золотом Стандарте. Полное совпадение по одному из учитываемых параметров (лемма, часть речи, грамматические признаки) автоматически получало оценку 0. При этом на дорожках без дизамбигуации для признания ответа правильным достаточно было наличия правильного разбора среди любого количества вариантов разбора, предложенных системой.

Случаи расхождений отправлялись на рассмотрение экспертам, которые должны были оценить их по следующей шкале:
1 – права Система;
2 – прав Золотой Стандарт;
3 – спорный грамматический вопрос;
4 – затрудняюсь определить (такие оценки впоследствии пересматривались в более широком кругу экспертов);
5 – неправы оба – и Система, и Стандарт.

Сравнение ответов систем с Золотым Стандартом позволило выделить наиболее распространенные отклонения от разборов, признанных эталонными.

1. Существенную часть ошибок составляет неправильное распознавание нестандартных классов слов. Можно выделить 5 основных типов.
1.1. Слова, имеющие дефис в графической репрезентации. Многие парсеры последовательно разбивают такие слова на части и лемматизируют их по отдельности, что можно признать правильным лишь в небольшом количестве случаев. Правомерность такого разбиения зависит от статуса элементов, составляющих дефисную конструкцию. Так, первым элементом может быть префиксоид (*штаб-квартира*), первый сегмент заимствований, не несущий в русском языке смысловой нагрузки (*Тянь-Шаня, холд-ап*), неотделимая часть некоторых типов предлогов (*из-за*) и наречий (*по-птичьи*) и т. д., и тогда подобное решение грамматически некорректно. Разбиение наиболее правомерно лишь тогда, когда обе части такого формального слова склоняются (например, когда одна из них является приложением к другой: *шофер–предприниматель*) и первая часть может обладать самостоятельными грамматическими признаками, но эти случаи составляют незначительную долю всех слов с дефисами.
1.2. Некоторые имена собственные. Неверно распознаются и лемматизируются по исходному сегменту. Проблемы частеречной принадлежности и грамматических признаков возникают не только с экзотическими словами, но и с фамилиями на *–ов*, *–их* и т.п.
1.3. Аббревиатуры. В отдельных случаях не распознаются вообще, некоторые системы опознают только часть речи, в той или иной мере – грамматические признаки.
1.4. Редкие слова. Зачастую также не распознаются или лемматизируются путем копирования сегмента исходного текста. Иногда по такой неправильной лемме определяются грамматические признаки.
1.5. Общепринятые сокращения типа *тыс.*, *ст.* («статья») и др.

Таким образом, большая часть ошибок возникает в «несловарных» словах, что объясняется тем, что парсеры либо имеют недостаточно эффективные средства обработки таких слов, либо вовсе их не имеют, полагаясь на закрытый список, составляющий словарь системы. Обилие ошибок с определением части речи и грамматической характеристики таких слов указывает на необходимость использования методов, учитывающих контекст. Экспертиза дорожки «Редкие слова» показала, что наиболее уязвимы для парсинга слова непродуктивных моделей (*джоулево, гильоше*), а также глагольные и наречные словоформы. Как кажется, это связано с тем, что для многих прикладных задач выбор в пользу продуктивных моделей и имен существительных дает большую эффективность системы.

2. Омонимия.
2.1. Достаточно типичными являются ошибки при разборе частичных (не «системных») омонимов, которые могли неверно лемматизироваться *(парный – парной)* и, как следствие, получали неверную POS–характеристику (*ели*).
2.2. Особый класс среди омонимов составляют пары из глаголов и отглагольных прилагательных/существительных (*окружающий* как форма глагола и как прилагательное, *данные* как форма глагола и как существительное), наречий и прилагательных (*ясно* как форма наречия или прилагательного), а также наречий и производных предлогов (*вблизи, навстречу*), для различения которых нельзя обойтись морфологическими критериями. Это обстоятельство вызвало некоторые колебания среди экспертов в оценке таких случаев.

3. Часть ошибок можно объяснить неправильным разбором по аналогии. Наиболее типичным случаем является ошибочная лемматизация глаголов с постфиксом *–ся* путем

отсечения этого постфикса в ситуации, когда соответствующий парный глагол не существует или отчетливо отличается по значению. Например, для глаголов типа *являться*, *стремиться*, *находиться* отдельными системами были предложены в качестве лемм, соответственно, *являть, стремить, находить*.

4. В отдельных случаях участники использовали классификации частей речи, которые не совпадали с предварительно заданной для данного соревнования, а потому использование символов этих классификаций оценивалось как ошибочное. Вместе с тем, по общей договоренности, исключение было сделано для числительных и местоимений, разбор которых не входил в зачет.

Наряду с вышеперечисленными типовыми ошибками был выделен ряд случаев лемматизации, определения части речи и полного грамматического разбора, которые по общему мнению были квалифицированы как спорные (оценка 3) и допускали более одного правильного (не наказываемого штрафными баллами) варианта. Основные спорные грамматические вопросы включали:
1) определение леммы сравнительных и превосходных степеней наречий и прилагательных (показатель степени может сохраняться в лемме, или же может быть использована лемма положительной степени[3]);
2) определение леммы краткой формы прилагательного (лемматизация по полной / краткой форме);
3) определение леммы парных по виду глаголов (лемматизация по несовершенному виду / по совершенному виду / по тому виду, который присутствует в исходной словоформе);
4) определение леммы глагольных словоформ с постфиксом *–ся* (лемматизация с сохранением постфикса / без него[4]).

8. Результаты соревнования

В основу ранжирования ответов систем положены три базовые величины:
- n, общее количество ответов на дорожке – принято за константу для всех систем и соответствует числу словоформ, получивших разметку в Золотом Стандарте и входящих в зачет в соответствии с регламентом;
- f, количество неправильных ответов системы на дорожке: неправильными считаются ответы, получившие оценку экспертов 2 и 5 (см. выше п. 7);
- t, количество правильных ответов системы на дорожке: правильными считаются ответы, получившие оценку 0, 1, 3 и 4.

Организаторы форума не могли уступить искушению использовать такие популярные метрики качества функционирования лингвистических информационных систем, как точность и полноту. Вместе с тем при более внимательном рассмотрении выяснилось, что эти метрики могут быть использованы лишь в весьма усеченном виде, по крайней мере на начальном этапе существования форума, когда все процедуры, в том числе оценочные, только отрабатываются.

Это несоответствие связано с принципиальными отличиями в функциональной архитектуре между информационным поиском, из которого берут начало точность и

[3] Во втором случае формы наречий должны быть приведены к наречиям, а формы прилагательных к прилагательным.
[4] В последнем случае имеется в виду страдательный залог невозвратного глагола. Варианты лемматизации признаются равноправными за исключением тех случаев, когда глагол не употребляется без *-ся* (*удаваться* - **удавать*) или же значение глагола без *–ся* принципиальным образом отличается от значения возвратного глагола (*находить – находиться*)

полнота, и морфологическим парсингом. В ситуации оценки информационного поиска все пространство используемой коллекции документов делится на четыре области:

- t_p – документы, признанные релевантными и найденные тестируемой системой,
- f_n – документы, признанные релевантными и не найденные тестируемой системой,
- f_p – документы, не признанные релевантными, но найденные системой,
- $(n – (t_p + f_n + f_p))$ – все остальные документы,

что позволяет определить точность Precision как отношение $t_p/(t_p+f_p)$, а полноту Recall как отношение $t_p/(t_p+f_n)$ и дать этим величинам вполне осмысленную интерпретацию. Однако эта ситуация не находит прямых соответствий в морфологическом анализе текста. Если принять за единицу подсчетов словоформу (а не, допустим, отдельный тег или вариант разбора), то пространство размеченной коллекции текстовых словоформ будет разделено на три области:

- t_p – словоформы, оценка которых учитывается при ответах системы и для которых система дала правильный ответ (= t),
- f_p – словоформы, оценка которых учитывается при ответах системы и для которых система дала неправильный ответ (= f),
- f_n – словоформы, оценка которых учитывается при ответах системы и для которых система не дала ответа (= n – t – f).

Если разбираемый текст содержит словоформы, разбор которых по общей договоренности не подвергается оценке (как местоимения и числительные на данном форуме), случаи их окказионального разбора отдельными системами никак не могут повлиять на оценку этих систем, поскольку остальные участники изначально отказались от их разбора и общее основание для сопоставления результатов всех участников отсутствует. Если одной словоформе из Золотого Стандарта в ответе системы соответствует две словоформы с собственными разборами (например, *бело–кремовое* VS *бело* и *кремовое*), то они получают одну общую оценку. Таким образом, сумма $t_p + f_n + f_p$ является константой (n), обозначающей число словоформ, по которым предполагается давать оценку системе, пользуясь данной версией Золотого Стандарта (это справедливо для всех дорожек – с дизамбигуацией и без дизамбигуации).

Механический перенос формул информационного поиска

$$\text{Precision} = t_p/(t_p+f_p)$$

и

$$\text{Recall} = t_p/(t_p+f_n)$$

в данную область дает лишь частичный эффект: точность вполне осмысленно характеризует ту пропорцию ответов системы, которой можно доверять, тогда как полнота едва ли может получить разумную интерпретацию. Причиной этому является отсутствие каких–либо общих содержательных признаков для двух слагаемых в знаменателе формулы – числом правильных ответов системы t_p и числом случаев, когда система по ошибке не дала никакого ответа f_n (заметим, что в информационном поиске сумма t_p+f_n давала не что иное, как количество документов, считающихся релевантными для данного запроса). Деление числа правильных ответов на сумму разнородных слагаемых не поддается осмыслению.

Вместе с тем, есть возможность воспользоваться еще одной метрикой, заимствованной из информационного поиска, которой является «аккуратность»:

$$\text{Accuracy} = t_p / (t_p + f_n + f_p)$$

В связи с особенностью нашего выбора базовых величин для расчетов (n, f и t) эта метрика имеет вид:

$$\text{Accuracy} = t_p / (t_p + f_n + f_p) = t / n$$

и легко интерпретируется как общая оценка качества работы парсера, поскольку позволяет судить о том, какая доля словоформ получит правильный разбор данным парсером.

Существуют иные подходы к определению полноты и точности, см., например, Paroubek 2007: 111–112, где описаны возможные интерпретации этих понятий специально для морфологического анализа без дизамбигуации. При этом либо рассматривается ситуация, допускающая множественность разборов в Золотом Стандарте, что является нетипичным в нашем случае, либо, при сравнении с Золотым Стандартом, приписывающим единственно возможную интерпретацию, полноту предлагается определять так, как у нас определена аккуратность, а точность включает понижающий коэффициент за неразрешенную неоднозначность. Однако мы сочли удобным использовать при экспертизе дорожек 2010 года описанную выше единообразную трактовку метрики для обоих вариантов разбора: как с дизамбигуацией, так и без дизамбигуации.

«Дизамбигуация: Леммы»					«Дизамбигуация: POS»					«Редкие слова»				
Участник	t	нет	f	Accur.	Участник	t	нет	f	Accur.	Участник	t	нет	f	Accur.
Melon	2008	14	24	98.1%	Olive	1991	22	33	97.3%	Desert	59	3	13	78.7%
Peru	1970	1	75	96.3%	Pine	1991	5	50	97.3%	Beaver	53	8	14	70.7%
Chocolate	1964	43	39	96.0%	Cadet	1958	43	45	95.7%	Burlywood	52	4	19	69.3%
Turquoise	1934	75	37	94.5%	Maroon	1943	0	103	95.0%	Copper	47	4	24	62.7%
Timberwolf	1925	0	121	94.1%	Sherbert	1934	75	37	94.5%	Lavender	46	0	29	61.3%
Pink	1831	11	204	89.5%	Apricot	1769	11	266	86.5%	Shadow	42	0	33	56.0%
Sienna	1430	547	69	69.9%	Shamrock	1394	547	105	68.1%	Snow	10	63	2	13.3%
										Forest	3	70	2	4.0%
Всего ответов				2046					2046	Всего ответов				75
Медиана				94.5%					95.0%	Медиана				62.0%

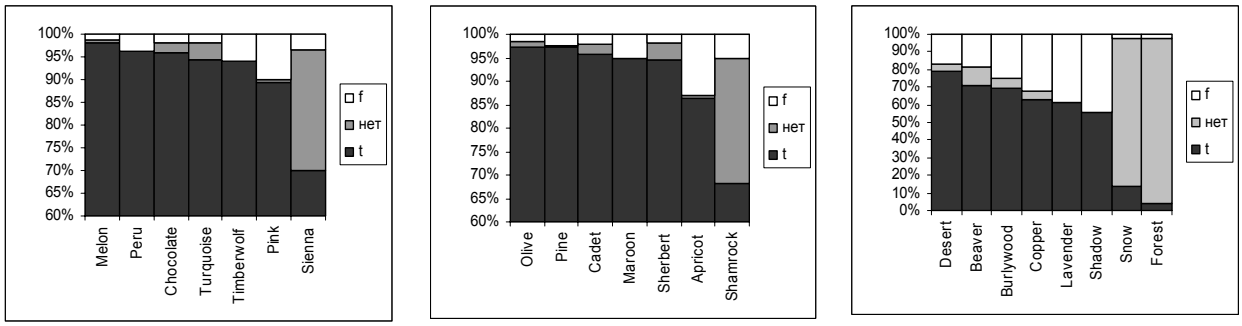

Таблица 1. Рейтинг систем на дорожках с дизамбигуацией и «Редкие слова».

9. Выводы, перспективы и задачи

Главной целью в 2010 году было положить начало проведению в России семинаров, посвященных оценке методов автоматического лингвистического анализа для русскоязычных коллекций. Как уже отмечалось, в мировой практике сложилась традиция проводить соревнования по различным аспектам автоматической обработки текста, в которых участвуют научные, научно–производственные, коммерческие разработчики, заинтересованные в независимой экспертизе. В России существует такая традиция в области информационного поиска (РОМИП). Однако соревнования, где основное внимание уделяется собственно лингвистическому анализу текста, в русскоязычном сообществе проводилось впервые.

В 2010 году был проведен комплекс работ, в результате которого удалось:
– апробировать организационные процедуры для такого рода соревнования и механизмы взаимодействия, в том числе дистанционного, в рамках оргкомитета;
– собрать большую коллекцию неразмеченных текстов разных жанров, на которой тестировалась работа систем;

– создать коллекцию Золотого Стандарта, размеченную вручную и выверенную несколькими экспертами; эта коллекция может быть использована в дальнейшем для тестирования систем и при подготовке специалистов по прикладной лингвистике;
– выработать основные принципы морфологической разметки для создания Золотого Стандарта;
– принять основные грамматические решения, обеспечивающие унификацию оценки разметок систем;
– выявить сложные и спорные случаи морфологической аннотации, вызывающие затруднения не только при автоматическом анализе, но и при разметке экспертами;
– провести оценку работы парсеров по четырем дорожкам для систем без дизамбигуации и по двум для систем с дизамбигуацией;
– провести содержательный анализ ошибок парсеров, выработать классификацию ошибок систем, а также решений, альтернативных принятым в Золотом Стандарте;
– анализ результатов выявил также сложности в применении к оценке морфологического анализа традиционных метрик, используемых в оценке информационного поиска.

В силу принципиальной несводимости к единому стандарту решений отдельно взятых систем по отношению к ряду спорных вопросов русской морфологии, в 2010 году эти спорные вопросы были вынесены за рамки соревнования. В дальнейшем предполагается постепенно сужать их круг и расширять лингвистическую базу для проведения соревнования, опираясь на взаимодействие с разработчиками морфологических парсеров и учитывая новейшие тенденции в этой области.

Как и ожидалось, анализ результатов работы систем морфологического анализа выявил целый ряд дискуссионных аспектов технологий морфологического анализа:
– состав набора морфологических тегов (специфика категоризации частей речи для различных задач);
– оптимальные соотношения между размером словаря и мощностью генератора гипотез для «несловарных» слов;
– способы борьбы с различными типами «системной» омонимии и др.

Были решены главные задачи форума 2010 года: построение типологии проблем автоматического морфологического анализа текста и оптимизация структурирования соответствующего набора данных, что в целом может служить дополнительным стимулом развития алгоритмов в этой области. Активное участие в соревновании большого количества различных научных и коммерческих коллективов в 2010 г. показало актуальность и востребованность проведения подобных форумов. Проявленный к форуму интерес укрепил уверенность в том, что этот проект положит начало ежегодным соревнованиям, целью которых является оценка методов и алгоритмов лингвистического анализа разного уровня. Последующие мероприятия могут быть посвящены синтаксическому и семантическому анализу, фактографии, анализу звучащей речи, использованию лексикографических ресурсов и многим другим аспектам автоматического анализа текста.

Список литературы

1. Коваль С.А. О сравнимости и эквивалентности компьютерных представлений морфологии // Компьютерная лингвистика и интеллектуальные технологии. Тр. междунар. конференции Диалог'2003 (Протвино, 11–16 июня 2003 г.) / Под ред. И. М. Кобозевой, Н. И. Лауфер, В. П. Селегея. М.: Наука, 2003. С. 305–311.
2. Российский семинар по Оценке Методов Информационного Поиска. Труды РОМИП 2009 (Петрозаводск, 16 сентября 2009г.). Санкт-Петербург: НУ ЦСИ, 2009.

3. Paroubek P. On the evaluation of the automatic parsing of natural language // Evaluation of text and speech systems. Text, speech and language technology. Vol. 37. Springer, 2007. P. 99–113.

NLP EVALUATION: RUSSIAN MORPHOLOGICAL PARSERS

Astaf'eva I., Bonch-Osmolovskaya A., Garejshina A., Grishina Ju., D'jachkov V., Ionov M., Koroleva A., Kudrinsky M., Lityagina A., Luchina E., Sidorova E., Toldova S.,
Moscow State University,
Lyashevskaya O., Savchuk S., Institute of Russian Language RAS,
Koval' S. *(lingtecheval@yahoo.com)*

NLP Evaluation forum (http://ru-eval.ru) is a new initiative aimed at independent assessing the methods that are used in Russian-oriented linguistic resources. The paper describes the first contest of morphological parsers, its participants, data and test collections and reports the design and results of evaluation as well as problematic cases.