Международная конференция «Диалог 2010»

Круглый стол «Извлечение знаний: проблемы и реальные успехи»

Ведущий В.Ф. Хорошевский (Вычислительный центр им. А.А. Дородницына РАН)

В.П. СЕЛЕГЕЙ (Вступительное слово от Программного Комитета): Когда речь на Диалоге идет об извлечении знаний, всегда чувствуются какие-то элементы рекламы, когда мы слышим соответствующие доклады. Что-то делается, но мы редко видим evaluation, то есть не очень понятно, на какой лингвистической, онтологической базе на самом деле работают системы, насколько они масштабируемы, как легко переходить от одной задачи к другой. Может быть, это и понятно узкому кругу лиц, причастных этой области, но очевидно, что это не очень понятно аудитории. Мы давно хотели обсудить вот эту большую и важную тему, с которой, тем не менее, все не очень ясно даже на уровне названия. Вот у нас тут указано «знание», и существует, например, обширная литература о том, чем «знание» отличается от «информации». Но пока есть множество похожих работ, позиционируемых как извлечение информации (IR- Information Retrieval), извлечение знаний, data mining и т.п. Хотелось бы разобраться, чем отличаются эти области, и какая за всем этим стоит наука, какое «тайное знание», какие реально используются модели, лингвистические, онтологические и прочие.

И тут нам так повезло, что за это дело согласился взяться Владимир Федорович Хорошевский. Он отдалился от «Диалога» на целое десятилетие и, вдруг, прислал в этом году доклад, который вы сегодня услышали. И стало понятно, что Владимир Федорович следил все же за тем, что происходит на «Диалоге» в этой области, и даже это исследовал. И мы ему очень благодарны и за это исследование и за то, что он согласился вести этот круглый стол, который, надеюсь, пройдет по гамбургскому счету.

В.Ф. ХОРОШЕВСКИЙ: Коллеги, наш круглый стол называется «Извлечение знаний: проблемы и реальные успехи». Я не случайно вынес на заставку [слайда] ту же картинку и надпись «Памяти Саши Нариньяни». Поскольку прошло совсем немного времени, как один из основателей «Диалога» и человек, который вложил в него очень много и души, и сил, и знаний, и всего остального, ушел из жизни. Прошел ровно месяц. Давайте будем обсуждать так, как будто бы он с нами.

[О регламенте, представление участников Круглого стола]

Прежде всего несколько слов о предмете обсуждения. Я бы хотел поточнее сформулировать тему обсуждения. Что мы, собственно, обсуждаем? Мы обсуждаем тему information extraction или мы обсуждаем тему knowledge acquisition?

Русский язык вообще очень активно сопротивляется терминам, в отличие от английского языка, где очень легко образуется новая терминология. Тем не менее мне кажется, что водораздел проходит вот по каким составляющим. Когда мы говорим о knowledge acquisition, то, вообще говоря, для нас главным все-таки является направление искусственного интеллекта и методы и средства искусственного интеллекта в приложении к задаче приобретения знаний. Когда же мы говорим об извлечении информации из текстов, здесь все-таки главным является лингвистика — вычислительная лингвистика, компьютерная лингвистика, которая в частности использует и методы искусственного интеллекта.

Поэтому, вообще говоря, я хотел бы сосредоточиться на information extraction. И что касается подходов, методов и инструментов, мне кажется, что на сегодняшний день ситуация такая, что существует, грубо говоря, три основных подхода, если не считать словарный подход, который мне лично не очень нравится, хотя он дает хорошие результаты. Это, так сказать, статистический, shallow подход и deep подход. В каждом из этих подходов имеются свои достижения и свои, так сказать, печали. Но гибридный подход на сегодняшний день представляется наиболее интересным, с моей точки зрения.

Что касается методов разработки систем типа information extraction, то здесь наблюдается следующее. Ну, классический анализ, с точки зрения лингвистики классический, анализ корпусов, причем представительный, — это тяжелый труд, для этого нужны средства автоматизации, и, возможно, мы еще про это успеем поговорить. Дальше — ручная спецификация правил шаблонов, скажем так, — это опять труд квалифицированных лингвистов-экспертов, которые к тому же почти сидят на двух стульях. Программисты и лингвисты, в «Avicomp’е» есть даже такая должность — лингвистический программист, скажем так. И, наконец, последний метод, который, к сожалению, пока очень мало используется, это машинное обучение с целью, например, генерации правильных шаблонов. И это мне кажется очень перспективно.

Что касается инструментов реализации, мне кажется, что на сегодняшний день существуют следующие вещи. С одной стороны, это программирование на обычных языках программирования — C, C++, Java и масса других языков. Понятно, что язык низкого уровня для таких задач, но зато позволяет выжать эффективность до тех пределов, до каких это возможно. Следующий шаг — это программирование на так называемых скриптовых языках, или языках символьной обработки — тут и всяческие Pearl’ы, PHP, Prolog’и, Lisp’ы, Рефалы и так далее. Наконец, следующий этап в развитии — это программирование на специальных языках представления знаний. И мне этот путь кажется наиболее перспективным, поскольку на сегодняшний день в программном обеспечении освоены методы эффективной трансляции с таких языков, а следовательно, можно получить хороший баланс между уровнем языка и эффективностью рабочей программы. Здесь всевозможные HMML и другие языки правил.

И, наконец, что мне кажется очень важным и что в нашей действительности очень редко по-настоящему используется, это так называемые FrameWork’и, или

платформы инженерии текста. Одна из первых таких платформ, которую вы, по-видимому, знаете, это платформа GATE, разработанная под идейным руководством Йорика Уилкса в Шеффилдском университете Великобритании. Сейчас там работает довольно большая команда, и лидер ее — Hamish Cunningham, и они активно работают над развитием именно платформы. Альтернативная платформа — это UIMA, которую предложила фирма IBM. И здесь идея была, так сказать, перекрыть весь спектр не только инженерии текстов, но и инженерии видеообразов, звуковых рядов и так далее. Что касается текста, то Богураев, который является руководителем этого проекта в IBM, — большой друг Каннигэма, и они попытались слиться в экстазе с GATE'ом, но, насколько я понимаю, из этого мало чего получилось.

Вот что касается подходов, методов и инструментов.

Анализ текущей ситуации. На слайде перечислены некоторые из российских организаций, которые, на мой взгляд, в той или иной мере вложили свой вклад в эту тематику. Что касается остального мира, то, к сожалению, ситуация вообще тяжелая, и писать обзоры очень тяжело. Приведу свой обзор в журнале «Искусственный интеллект и принятие решений», № 1 за 2008-й год. Ну а также надо смотреть труды всех конференций, TREC'и, MUC'и, DUC'и, LREC'и и так далее.

Реальные результаты — это то, что, на мой взгляд, по гамбургскому счету. Умеют уже все: извлечение некоторых для named entity, person, organization, location — маленькие entity, которые извлекаются на уровне регулярных выражений, PostAddress'ы, и, пожалуй, всё. Что касается извлечения отношений между объектами, то в малый джентльменский набор входят EmployeeOf и LocatedIn. Правда, мне сказали, что еще одним освоенным отношением является «они говорят», «о них говорят». Умеют далеко не все, хотя многие говорят, что умеют, извлекать расширенные типы NE. Когда мы говорим о расширенном типе NE, это и продукты, это и многое другое. Причем я говорю не о том, что извлекать по словарю. Это умеют все, это неинтересно, потому что словарь вы никогда не запомните целиком в силу появления все новых и новых примеров этих объектов. Определять семантические отношения широкого спектра, которые характеризуют, вообще говоря, некоторые факты и события, и, так сказать, формировать time-line'ы — вот это на самом деле одна из трудных и интересных задач, когда вы из текста извлекаете историю во времени развертывания событий. Ну и, наконец, не умеет почти никто или никто — это всерьез обрабатывать анафору и эллипсисы, хотя многие утверждают, что они с этим справляются, — неправда, не верьте. Определять тональности — все говорят, что это одна из самых интересных задач, и «Медиалогия» вроде бы все решила. На самом деле мы знаем, как этот вопрос решила «Медиалогия». А что касается тональности по отношению к чему-то, это гораздо более сложная задача. А уж если в конце будет стоять «Не правда ли?», а до этого была хвалебная статья, которая будет полностью перечеркнута последней фразой, вот этого не умеет делать никто. И, наконец, последнее, что почти никто не умеет, это формировать картину мира на основании коллекций. Когда мы говорим об обработке одного текста, всё нормально с точностью до

наших умений и знаний. А вот когда у вас есть десять миллионов текстов, и вам надо на основе этой коллекции построить картину мира, всё получается значительно сложнее. Нерешенная проблема — вот, на мой взгляд, это бесшовная интеграция лингвистических и предметных знаний, интеграция всего спектра, разработка методов оценивания и сравнения, evaluation, вот одна из самых больных задач, ну и, наконец, полномасштабное изучение информации, использование методов машинного обучения. А с точки зрения прикладной — это интеграция вот этих подсистем как компонент в более сложные системы.

И тенденции развития. Я их на слайде перечислил, я полагаю, что вы читаете быстрее, чем я говорю. Спасибо за внимание.

Н.В. ЛУКАШЕВИЧ (НИВЦ МГУ им. М.В. Ломоносова): Я хочу обсудить несколько пунктов из перечисленной повестки. Сначала я хотела высказаться по поводу information extraction и knowledge acquisition. Действительно, тематика извлечения знаний идет из искусственного интеллекта, и как раз обсуждалась проблема, что откуда-то нужно брать знания и помимо общения с экспертами, в частности можно извлекать такие знания из текстов. Мне кажется, что различие между этими терминами лежит еще дополнительно в том, что тематика knowledge acquisition, то есть все-таки больше извлечение знаний, лежит такая тематика, которая сейчас, например, называется onthology learning, то есть обучение онтологии, это как извлечение терминов, то есть это наполнение таких знаний о типах сущностей и о типах их отношений. В то время как information extraction все-таки больше тяготеет к извлечению конкретных сущностей и отношения между ними. То есть у меня картинка именно такая: если мы хотим автоматически извлекать термины из текста, то мы ближе к области knowledge acquisition, если мы извлекаем именно сущности, то это information extraction.

В.П. СЕЛЕГЕЙ: А можно привести пример?

Н.В. ЛУКАШЕВИЧ: Я же сказала, что если мы извлекаем термин из текста, для меня это извлечение знаний, knowledge acquisition. Новые, неизвестные нам термины. Мы вот пришли в новую область, и мы хотим узнать, какие там термины.

В.Ф. ХОРОШЕВСКИЙ: Ну вот, скажем, «появился iphone такой-то» — это термин.

Н.В. ЛУКАШЕВИЧ: Дискуссия — не сейчас. Это, во-первых, не термин… Теперь хотелось бы поговорить о состоянии, собственно говоря, information extraction, то есть извлечения информации, которая представляется более закрытой областью в России. Какие достижения имеются в извлечении терминов, это можно

прочитать в работах российских авторов. Но мы, к сожалению, не можем из наших российских работ узнать ничего о том, какой уровень современного достижения в извлечении именованных сущностей, хотя Владимир Федорович нам так эффектно сказал, что куча проблем решена, а что-то не решено, на самом деле, если кто-то хочет узнать — решено до какой степени и не решено до какой степени, то практически по российским работам это узнать невозможно. Я лично знаю, что… может быть, я ошибаюсь, есть одна работа Александра Ермакова, опубликованная в журнале «Информационные технологии», по очень специфической задаче — извлечению оценочных суждений об автомобилях из блогов, где он четко приводит цифры, как же ему удалось, то есть точность и полноту, с которой ему удалось это сделать. Точность в его подходе была 80%, а полнота — 20. На самом деле, таких честных цифр, которые хоть в какой-то мере ставят ситуацию в какое-то понимание, что на самом деле настолько много еще нужно делать, — очень этого не хватает. Сегодня Владимир Федорович сказал в своем докладе, что референцию разрешает неплохо. Вот прежде всего мы слышим отовсюду это «неплохо», но в России картину в цифрах узнать абсолютно невозможно. Здесь как раз хочется обратиться к зарубежному опыту и вспомнить конференцию «Message Understanding Conference», которая началась в 1991 году и как раз начали исследовать, как же извлекаются разные типы информации из текстов. И как раз организаторы этой конференции писали: «Мы организуем вам эту конференцию затем, чтобы хотя бы понять, какой уровень достижений». Собственно говоря, что могут и что не могут делать системы. То есть эта проблема довольно общая, если за границей с 1992 года цифры таким образом стали достаточно открыты и известны, то в России как-то не удается ни в рамках РОМИПа, ни каких-то отдельных увидеть такие какие-то соревнования или цифры в отдельных работах. Соответственно, только читая западные материалы конференций, можно узнать, что именованные сущности, вот такие неизвестные, извлекаются 90—95%, что какой-то набор отношений более-менее произвольный — около 70%, я имею в виду, то есть сочетание точности и полноты, а уже какие-то сложные ситуации и факты — 50—60%, а реальная кореференция — тоже порядка 50—60%. И отсюда как раз видно, что не все так прекрасно.

Последний пункт, который я хотела бы сказать, это о проблемах. У Владимира Федоровича услышала, что некий большой потенциал он видит в применении машинного обучения. Это, безусловно, так, но не в чистом виде. В чистом виде это означает следующее: некто или какая-то группа людей сидит размечает эти сущности и отношения между ними, тратит безумное количество времени сначала, чтобы разметить, а затем программы машинного обучения должны обучиться этому. Мне кажется, что это ни чем не лучше, чем делать какие-то шаблоны и писать некие правила. И поэтому, в чем я соглашусь, это что сейчас такое важное направление — это гибридный подход, связанный с тем, что как бы надо нам описать небольшое количество шаблонов, ручной делать работы, и затем автоматизированно попытаться, так сказать, нарастить, методами машинного обучения в частности, помочь человеку нарастить эту базу шаблонов. Потому что известно, что эти системы, как называлось deep aproach, или на шаблонах работающие, то есть то, что пишут люди, эти системы характеризуются

очень высокой точностью и достаточно низкой полнотой. Это и проблема того, что Владимир Федорович назвал глубоким подходом, и shallow подхода на шаблонах. Если это пишут люди, они пишут точно, но неполно.

Д.В. ЛАНДЭ (Информационный центр «ЭЛВИСТИ»): Тема действительно очень сложная. До сих пор я не сильно воспринимаю терминологию, связанную со знанием, особенно с их изучением, хранением, сохранением. Почему — я считаю, что знания — вещь субъективная, в принципе извлекать надо информацию. Это такой философский вопрос. Я понимаю, что знания лежат в голове человека, который чему-то научился, воспринял информацию и дальше уже ее приспособить, дальше уже ее сохранение и так далее, с точки зрения даже философии и с моей точки зрения, теряет смысл. Хотя, может быть, я чем-то обижу тех, кто десятилетиями, двадцатилетиями, тридцатилетиями занимаются этим направлением. Вместе с тем у нас появилось много новых хороших лозунгов, слоганов, особенно мне нравится information extraction — с моей точки зрения, это только кусок, только составная часть… На самом деле вроде как поменяли название, и сразу может меняться…

В.Ф. ХОРОШЕВСКИЙ: Финансирование.

Д.В. ЛАНДЭ: Нет, скажем так. Если говорить, что некий термин искусственного интеллекта набил чиновникам, которые хотели получить быстрый результат, оскомину, то, наверно, иногда надо менять терминологию, мы никуда не денемся.

В.Ф. ХОРОШЕВСКИЙ: Ну американцы так сплошь и рядом делают.

Д.В. ЛАНДЭ: Text mining почему-то включает в себя кластеризацию, классификацию, информационный поиск и feature extraction.

Теперь, подходы, методы и инструменты. Действительно, извлечение именованных сущностей — задача очень конкретная. И так с разбегу говорить, что она решена большинством из тех, кто за нее брался, — это плохо. На самом деле там есть очень много подводных камней, и я считаю, что качественное извлечение даже именованной сущности без лингвистов, с помощью одной статистики, можно сделать до определенного уровня. Конкретный пример: очень хороший у меня есть инструмент, который извлекает имена из текстов — из ровных текстов, повествовательных предложений, всё хорошо. Это же дело — подставляется база данных резюме людей, которые ищут работу, заодно они пишут, где они работали и как связаны их имена с тем, где они работали, можно определить… Самое распространенное имя оказалось — «пол женский». По всей

видимости, статистика без понимания смысла, без понимания, где она применяется, работать не может. И тут я не согласен с одним тезисом Владимира Федоровича, что статистические методы базируются на обработке больших корпусов. Они могут быть рядом. Ну вот пример той же системы по извлечению знаний, Autonomy. Они работают на огромном количестве языков, на языках, которые пишутся слева направо, сверху вниз, все равно. Они смотрят на всю нашу лингвистику как на последовательность символов с разделителем. При этом прекрасные результаты получаются.

В.Ф. ХОРОШЕВСКИЙ: Ну как вам сказать. Когда я разговаривал с руководителем подразделения Autonomy, то я им предложил простую задачу про Пушкинскую площадь. И Autonomy на этой фразе, хотя они обрабатывают русский язык, рухнула моментально. Она Пушкина интерпретировала совершенно не как человека. Площадь Пушкина.

Д.В. ЛАНДЭ: Совершенно коммерчески успешная система. Но мы с вами понимаем, что таким путем в чистом виде идти нельзя. Может быть, коммерческий успех повторить можно, но идти нельзя. Надо, в общем-то, использовать и корпусный метод как наиболее простой способ, потому что все-таки синтаксический, семантический разбор предложения — вещь весьма сложная, это один из самых главных предметов прикладной лингвистики на сегодняшний день. Задача не решена для украинского языка, наверно, для русского языка тоже. Про ситуацию «Россия против всего мира» рассказывать, наверно, тяжело. Наталья правильно сказала, и мне приходится соглашаться и самому — очень плохие дела с оценками тех систем, которые сделаны в России, и тех систем, которые работают. Мы можем их оценить только по тому состоянию, насколько довольны наши пользователи или редкие экспертные группы. Я видел формулы, там нолики-единички суммируются и так далее, ну это мнение экспертов, которые показывают выборные технологии, например избирательные технологии совсем неправильно. Собственно говоря, это самая наша нерешенная проблема. Теперь если говорить о применении тех светлых идей, которые есть, без кавычек нормальных идей, для технологии как извлечения информации, — они стоят перед нашими лидерами в том числе, то есть перед RCO, который считается лидером в России, наверное, это правильно, — это проблема быстрого действия. Всё работает очень медленно. Может быть, эта проблема будет решена.

Тенденции развития. Перспективы имеются светлые. По крайней мере проблему быстродействия мы решим. Проблема синтаксического разбора — я думаю, что, в принципе, были реализованы очень большие шаги, то есть мы увидели реальные алгоритмы. Соотношения полноты и точности — эти соотношения надо получить, это раз. Во-вторых, понятное дело, что действительно точность на контрольных всяких примерах получается очень хорошая.

В.Ф. ХОРОШЕВСКИЙ: Особенно на своих.

Д.В. ЛАНДЭ: Да, особенно на своих. Ну и методология не допускает, конечно, таких вещей… Я хочу сказать, что есть такой взгляд, он имеет право на жизнь, что полнота при развитии современного информационного пространства в общем-то и не нужна. Можно поспорить.

Е.И. БОЛЬШАКОВА (МГУ им. М.В. Ломоносова): Хочу сказать, что все-таки даже с нашими рассуждениями о полноте и точности, даже если все опубликуют свои цифры, то это еще не гамбургский счет. А гамбургский счет начинается тогда, когда все приносят свои программы и начинают их демонстрировать, измерять относительно определенной технологии, относительно определенной задачи, как это делается на некоторой серии конференций. Вопрос — готовы ли мы к этому? В нашей июньской беседе прошлого года вы сказали, что мы не готовы. С другой стороны, завтрашний день интересен тем, что нам покажут сравнение морфологических парсеров, это большой шаг вперед. Вопрос — следует ли нам стремиться к этому? Ну я предлагаю для начала все-таки действительно, поддерживая предыдущего докладчика, все-таки опубликовать реальные цифры. Потому что мой опыт показывает… на самом деле я относительно молодой участник из панелистов, я немного даже удивлена, что меня сюда пригласили, потому что в отличие от Натальи Валентиновны, которая всю свою научную жизнь занимается извлечением знаний, я всего лишь лет 12 этим занимаюсь. На самом деле у меня было довольно долгое вхождение в эту область, и здесь требуется определенный лингвистический background, и его было тяжело получить, но его было легко получить именно для русского языка благодаря нашей лингвистической ветви. Но тогда если брать вид именно computer science, computational linguistics, то здесь ситуация, конечно же, не очень хорошая. В каком смысле? Я честно пыталась обозреть всю область, попыталась даже написать, может быть, какой-то обзор — ситуация такая, что даже если люди чем-то занимаются, они почему-то про это не пишут, а если пишут, то совершенно непонятно, не раскрывают свои методы. Это еще вопрос — как принимаются работы, в том числе на «Диалог». И это еще вопрос к Владимиру Федоровичу, который по своей довольно-таки мощной программе — у вас очень мало публикаций, и в том числе на «Диалоге». Для того чтобы нам достичь какого-то западного уровня, надо по крайней мере публиковаться. Здесь очень сложный вопрос — если человек уходит в какую-то фирму, которая финансируется, то почему-то перестают быть публикации. Но тем не менее они должны быть, и на западе этот вопрос как-то решается, у нас, к сожалению, нет. Я знаю, что у нас фирмы, занимающиеся поисковыми системами, этим занимаются, но от этих фирм почти нет публикаций. Вопросы — как нам сравниваться, что нам обсуждать, куда нам двигать, и вообще, чтобы было творческое развитие, должно быть много идей. Выживает сильнейшая идея, если мало — совсем тяжело жить.

Интересный вопрос, вы говорите, — shallow и deep, то есть поверхностный и глубинный. Да, первоначально начали поверхностный, потому что глубинный — тяжело, много разборов… Но посмотреть на то, как, например, в реально работающих системах это shallow работает и сколько к нему наращивается примочек, мне кажется, может быть, с самого начала? И, кстати, так делают многие на западе. То есть они изучают information extraction начиная с того, что мелкая задача решается с того, что получил синтаксическое дерево разборов, а там уже начинают чего-то извлекать. Довольно симптоматично для них. Очень много приплюсовывается еще знаний энциклопедических, но опять же об этом плохо пишут.

И последний вопрос, на котором я хотела бы остановиться, касается вашего доклада относительно провала публикаций последних лет, касающихся платформ, инженерии, настраивания и прочего. Да, такие работы были, они были такого детского типа, но давно, их сейчас нет. К сожалению, вынуждена констатировать, что эти работы лежат на стыке software engineering и той проблематики, которая представлена здесь, то есть computational linguistics. Я не уверена, что подобные работы хорошо берутся на конференции, скорее всего плохо. Даже упомянутая система «Ellogon», я вообще удивляюсь, как она промылилась в лингвистическую. Это, в общем-то, software.

В.Ф. ХОРОШЕВСКИЙ: Это не только software, это тяжелый интерактивный software.

Е.И. БОЛЬШАКОВА: Да, и с ним тяжело. Хочу подчеркнуть одну мысль. Я бы приветствовала появление таких работ на этой конференции, хотя я не уверена, что это нужно сделать. У меня вопрос ко всем. Почему, потому что для того, чтобы проводить какие-то исследования, мы должны быстро настроить те же самые шаблоны, а не обеспечивать этим инструментарием сами себя. Наша группа фактически начала с того, что мы обеспечивали программным инструментарием сами себя, потому что — ну да, есть RCO, но нам не нравилось, что RCO это все-таки идейный слепок с GATE’а, и ничего тут не поделаешь. И почему надо выбрать всё, построенное по одной архитектуре? В общем, необязательно. Можно и в нашей стране заниматься этим, другое дело, что это не финансируется. Вот на этой интересной точке я закончу.

Л. ГЕРШЕНЗОН (Компания «Яндекс»): Я хотел совсем про другое говорить, я буду представителем бизнеса, индустрии, буду говорить совсем не как научный человек, не буду задавать вопросы, а буду давать ответы очень злые. Значит, я знаю, почему в России так плохо с information extraction и с соответствующими системами. Значит, почему так плохо с системной оценкой? Если мы вспоминаем американские MUC’и и DUC’и, мы должны вспомнить, откуда они пошли. Они

были инициированы американским военным агентством DARPA, если я не ошибаюсь.

В.Ф. ХОРОШЕВСКИЙ: Как и всё остальное.

Л. ГЕРШЕНЗОН: Как и всё остальное. Зачем нужны оценки, почему они там есть? Потому что заказчики — это прикладные вещи, прикладные системы. У них есть задачи, которые они не могут решить без соответствующих систем. Им нужно выбрать лучшее по соотношению разных качеств. Поэтому проводятся эти TREC'и, системы оценки, а участники заинтересованы в том, чтобы конкурировать и доказывать, что они лучше. Пока в России нет бизнеса, компаний, у которых какие-то серьезные процессы построены на системах извлечения знаний, text mining, fact extraction, knowledge acquisition и так далее, пока это остается такими игрушками. Как только появятся компании или люди, которые без этого не смогут жить, все будет с оценками, и с дорожками, и с TREC'ами. Такое мое мнение.

Я представляю компанию «Яндекс». Прошу прощения за непарадную форму одежду, я узнал о том, что буду участвовать в круглом столе, после того как вышел из дома. Кто не знает английского языка, у меня на майке написано приблизительно следующее: «Брось ты свой компьютер, давай выйдем на улицу, поговорим наконец». Написано «Fuck Google, ask me». Мы еще не сделали такое для «Яндекса», поэтому приходится рекламировать наших коллег и партнеров.

Для нас как для промышленной системы, для сервиса миллионного, важна постановка задачи. У нас довольно много, на наш взгляд, information extraction приложений. Почему information extraction в Яндексе? Потому что наша основная задача — поисковая. Мы столкнулись с тем, что поняли, что без более глубокого анализа, без выделения named entity, объектов и связей нам трудно развиваться. Как в какой-то момент появился морфологический поиск, так следующим естественным этапом стало появление объектов для использования их в поиске. Это задача номер раз. Вторая задача, для поисковых систем важная, это собственно представление результатов, традиционные снипиты. Если мы говорим про тенденции, то развитие основных поисковых систем в России и за рубежом связано с тем, что снипиты обогащаются, становятся красивее, структурированее, и это второе применение результатов таких систем. Третья история — это про то, что поиск и вообще поисковые задачи, поисковая модель пользователя меняется. Не всем, не всегда и не в основном нужно находить веб-страницы и тексты. Люди задают поисковым системам реальные вопросы. Если они ищут аптеку рядом с собой, чтобы купить там конкретное лекарство, их не интересует сайт данной аптеки, какой у нее page run и так далее. Таких вопросов все больше и больше. Поэтому системы information extraction решают проблемы реальных запросов из реального мира. И пример проекта «Пресс-портреты» — я его не рекламирую, но хочу отметить, мы его запустили 4 года назад, сегодня вышла новая версия, красивая, информативная, и всех приглашаю тоже поиграть-посмотреть — это

такой шажок в сторону специализированного, оторванного от интернет-страниц поиска. Это база людей и вся информация, которая в интернетах про людей собрана, — новый источник, пусть и вторичный, но отдельная база знаний. Тенденция — думаю, что в эту сторону, и эта тенденция началась не вчера, — разметка и создание новых специализированных поисковиков, которые, кстати, успешно на западе и в Америке конкурируют с нормальными поисковиками, это поисковики по объявлениям, по товарам, по вакансиям, по недвижимости и так далее.

Что хочется сказать про технологические вещи. Повторяю, что говорю как злой представитель индустрии. Есть выделение, определение анафорических связей, денотатов — нету, в промышленных системах такие вопросы… или нужно ли совмещать статистический подход с шаблонами, — эти решения вытекают из конкретных результатов. Грубо говоря, делается какая-то система на шаблонах, дальше мы видим конкретные проблемы, делается статистическая система, и вопрос гибрида и совмещения, на мой взгляд, возникает тогда, когда можно доказать-увидеть, что этот гибрид реально поможет решить конкретные проблемы. Наш опыт с пресс-портретами, хотя я хочу сказать сразу, что в Яндексе результатов fact extraction, information extraction довольно много, и все они заметны, «Пресс-портреты» — отдельный, целиком на этом построенный сервис, результаты майнинга адресов и показа их снипитов уже около года доступны со ссылкой на карты. Есть разные типы спец-снипитов — специальных… от рецептов до гостиниц… И есть результаты реально работающие — информация о каких-то named entity, информация о которых лежит в поисковом индексе, скажем, при запросе, содержащем ФИО в значении человека, эта информация используется при ранжировании. И дальше этого будет только больше. Я думаю действительно, что это актуальная тема — совмещение статистических и шаблонных подходов, потому что есть много примеров, с одной стороны, когда шаблоны упираются, а абсолютно стопроцентно подходящий под шаблон пример оказывается ошибочным, потому что просто не хватает знаний, и, скажем, человек понимает из более широкого контекста, что в «библиотека имени Ленина» — «Ленина» — это не про человека, а часть названия организации, и задача для informational extraction — использование более широкого контекста на уровне предложений, на уровне документа, на уровне, скажем, пользовательских сессий — я думаю, это задача на ближайшие какие-то годы.

Последнее, что я хочу еще сказать, есть два направления по извлечению информации. Владимир Федорович показывал слайды, и там везде было слово «аннотирование», направление именно создания какой-то новой базы знаний, где будет физический денотат. Грубо говоря, в «Пресс-портретах» это реальный человек. Вот его фотография, вот информация про него, вот профиль в социальных сетях, вот интервью и новости про него конкретно, значит, это направление будет развиваться. И здесь собственно две сложности — вопрос про аннотирование и вопрос про выделение объектов, связей и так далее с текстом. Есть история, про которую здесь не говорилось, про отождествление разных фактов для определения того, что эти два факта в разных текстах, в разных

документах, на разных сайтах — про физически один объект в реальном мире. Это очень сложная, интересная проблема.

Совсем последняя вещь. Когда мы говорим про извлечение знаний, у нас есть тоже два потока. Есть совсем естественные языковые тексты, их много — новостные сообщения, блоги, просто тексты, — и сильно или слабо, частично структурированные документы в Вебе, различные справочники, даже Википедия в значительной степени и так далее. Совмещение не просто статистических и лингвистических подходов, но и сливание информации из разных источников, — я думаю, эта тема в ближайшее время будет актуальна.

К вопросу о полноте и точности — что полнота — не тема. Тут я вспомнил такой анекдот про Шерлока Холмса, как с Ватсоном в шаре полетели куда-то, человек мимо проходил, спросили: «Где мы находимся?» Им ответили: «В корзине воздушного шара». Вот пример стопроцентной точности при достаточно низкой полноте. Вопросы полноты сейчас в поисковых системах получают новую значимость, новую важность. Довольно много задач, когда важно, даже в Интернете, ничего не пропустить, никакого ни факта, ни документа. Поэтому от них так просто отделаться не удастся. Спасибо.

Е.Б. КОЗЕРЕНКО (ИПИ РАН): Уважаемые коллеги, предыдущие докладчики сказали очень много того, что хотела сказать я. Чтобы не повторяться, я скажу то, что я хотела бы добавить. По первому вопросу, information extraction и knowledge acquisition, это верно. Если брать наше сообщество и весь остальной мир, мы имеем параллельно развивающиеся тенденции, два направления, причем между ними существует значительное пересечение, потому что когда-то доклады, с моей точки зрения, ближе к information extraction, попадают в конференцию, которая называется «Knowledge engineering», и, соответственно, один из разделов knowledge engineering — это knowledge acquisition. И наоборот. Из чего нужно сделать вывод, что весь остальной мир еще не определил, каковы между ними различия и как действительно это всё разграничить достаточно жестко, тем более методы и подходы для обоих задач и там и там бывают очень близки, а иногда и совпадают. Хотя, видимо, все-таки information extraction — это область действительно более, наверно, жестко задающая свои правила игры и все-таки ближе к задачам информационного поиска. И, наверно, information extraction — это более технологичная на сегодняшний момент и динамично развивающаяся область. То есть действительно уже сложили свои методы оценки результатов, и надежные методы, в отличие от такой более, наверно, еще пока расплывчатой и еще пока не очень определенной области, как knowledge acquisition. Что касается исследований, которые ведем мы, они все-таки ближе к направлению второму. Очень хорошо, что во вступительном слове Владимира Федоровича прозвучала историческая, обзорная нотка. Действительно, исследования в области инженерии знаний, искусственного интеллекта уходят корнями в амбициозные проекты, постановка задач которых, видимо, стимулировала не одно поколение исследователей, разработчиков, которые хотели сделать умную машину, которая

понимает человеческий язык, очень хорошо умеет извлекать знания и выдавать точные и правильные ответы. Пока на сегодняшний день это не так, хотя нельзя сказать, что ничего в этой области не сделано, и достаточно много. Просто уже стало ясно, как вот, скажем, область машинного перевода, задача не так проста, как она казалась в начале.

Что касается подходов, методов, инструментов и реальных результатов. Подходы есть, какие-то из подходов опять-таки достаточно традиционны, какие-то являются достаточно новыми. И в частности эти подходы связаны с таким бурным толчком развития методов стохастической парадигмы, того, что связано с машинным обучением и автоматизации процессов извлечения структур, того, что можно назвать структурированными знаниями, из данных. И в этом смысле два слова я просто обязана сказать про машинное обучение, потому что у нашего «Диалога» есть американский филиал, который называется «Intelligent and linguistic technology». Это такая небольшая конференция, workshop, в рамках конференции по искусственному интеллекту, которая ежегодно проходит в Соединенных Штатах как часть большого конгресса по computer science. На протяжении четырех лет эта конференция была включена в конференцию по машинному обучению. И это было очень полезно, потому что удалось понять суть того, что делается в этом направлении, и попытаться осознать, насколько это применимо и полезно для тех задач, которыми занимаемся мы. В двух словах — машинное обучение это, в общем, система методов и теоретических и инструментальных аппаратов, которые предназначены для того, чтобы автоматически строить модель некоторой области из объектов, относящихся к этой области. И вот машинное обучение может быть таким направленным, а может быть контролируемым, а может быть неконтролируемым. Что это значит? Если машинное обучение неконтролируемо, предполагается, что автомат без какой бы то ни было подсказки со стороны человека должен извлечь эту структуру данных и построить некую модель гипотезы. Для каких-то областей знаний, прикладных областей это возможно. Для естественного языка это тоже возможно, но до каких-то пределов. И, в общем-то, как сейчас показывает опыт и разработчиков, которые занимаются машинным обучением достаточно давно и успешно, тенденция к развитию таких supervised methods, там, где автомат обучается на заранее каком-то размеченном корпусе правильных примеров и далее может строить свои гипотезы, основываясь на этих данных. И возможно, как видится, для задач knowledge acquisition и в частности knowledge acquisition из текстов на естественном языке, видимо, такой supervised learning, контролируемое машинное обучение, наверное, это будет интересный подход, который даст, возможно, хорошие результаты, надо надеяться.

Нерешенных проблем много и, безусловно, это не только разрешение анафорических ссылок, а добротный анализ текста как такового. Всё, что мы имеем, это большее или меньшее приближение к какому-то ожидаемому результату. Причем, в отличие от результатов информационного поиска, не совсем понятно, как это можно считать с помощью какой-то программы. Я думаю, что уникальное средство оценки и сравнения — это то, что должны разрабатывать, причем эта задача, с моей точки зрения, не менее сложная, чем

разработка самих программ из области искусственного интеллекта и knowledge acquisition.

И что касается тенденции развития областей исследования, тут я всецело поддерживаю, наверно, точку зрения о гибридных подходах. Возвращаясь к нашему опыту и беседами со специалистами по компьютерному обучению, дело в том, что, скажем, такие алгоритмы на основе нейронных сетей, генетические алгоритмы тоже используются для задач машинного обучения. У нас на одной из конференций был американский коллега, который сделал несколько работ по применению искусственных нейронных сетей для анализа предложений, достаточно простых предложений естественного языка. И чтобы распознать простую фразу, которая с перевода на русский язык означает, что собака гналась за кошкой, там требовался очень громоздкий ход. Когда мы с ним попытались обсуждать, что нельзя ли в эту систему какие-то заранее заданные правила, мы же знаем, что это так, зачем системе это, как говорится, изучать, извлекать автоматически, если эти правила есть, они достоверны, и их можно использовать. Он подумал и через год приехал с докладом, посвященным методам моделирования предложений пропозициональной логики средствами нейронных сетей. Это возможно. Будут ли эти направления развиваться, увидим. Представляется, что это интересно.

А с другой стороны, включение статистических данных в какие-то правиловые системы, пусть это будут правила распознавания предложений естественного языка или правила для принятия решений при обработке извлеченных знаний, то есть применение правил, записанных на языке представления знаний, введение туда весов для того чтобы какая-то гипотеза могла оцениваться и приниматься системой при выводе и при принятии решений — наверно, это полезно и верно. Потому что выразить всё только правилами — это нереально и не надо, это переусложняет систему и не дает желаемого результата, то есть слишком много правил — это плохо.

В.П. СЕЛЕГЕЙ: Меня попросили рассказать о впечатлениях от последней конференции LREC в интересующей нас сфере. Мне кажется, цифры, которые я попробую привести, любопытны. Поскольку не все здесь присутствующие были в первый день конференции, я напомню одну цифру с LREC – конференции, задачей которой является анализ существующих лингвистических ресурсов. Раз в два года собираются люди и сообщают, какие ресурсы они сделали и для каких целей. В этом году было 1200 участников и было представлено 1200 ресурсов разного типа — от языковых данных моделей, алгоритмов, технологий – ведь все это может считаться ресурсами.

В первый день на «Диалоге» мы говорили о роли русского языка, дискуссия была специфическая — стоит ли «Диалогу» переходить на английский в качестве рабочего, сравнивались данные о русском языке на «Диалоге» и на этой конференции LREC. Выяснилось, что русский язык не входит даже в первые 20

наиболее часто используемых, обсуждаемых языков этой конференции. Там на первом месте был английский — примерно с тысячу цитирований, на 20-м месте был венгерский с 22 цитированиями, соответственно, русский язык еще меньше. Такая печальная цифра, которая заставляет о многом задуматься. При том, что уехало из России немало исследователей, но русским они не занимаются.

Ближе к теме теперь. Откуда берутся цифры? В этом году создатели LREC'а попытались сделать так называемую лингвистическую карту. Приезжают люди со всего мира, представляют свои ресурсы. И организаторы решили сделать такой информационный проект, обобщающий все эти ресурсы по языкам и задачам. Эти 1200 ресурсов были каталогизированы, посчитаны, и эти данные можно теперь увидеть, они опубликованы в Интернете. И что существенно, это ресурсы скорее академические, чем коммерческие, и они в среднем являются доступными, хотя есть и исключения. Я очень благодарен Льву за то, что он в ходе своего рассказа упомянул такую важную особенность — что все способы evaluation появляются тогда, когда появляются заказчики. Пока разработчики просто вешают лапшу на уши друг другу, все их методы оценки в общем-то мало чего стоят. И в области искусственного интеллекта вообще и извлечения знаний в частности, не будем скрывать, эта ситуация продолжалась десятилетиями, все это более-менее понимают. Но вот сегодня она изменилась, и важным свидетельством этого является как раз этот вот этот каталог LREC, а именно характеристика назначения ресурса по мысли его разработчиков. Мы можем увидеть, о каких задачах прежде всего думают люди, когда они создают лингвистические ресурсы. Если взять за 100% все назначения, все цели, которые были указаны разработчиками ресурсов, выясняется очень интересная вещь: на первом месте в качестве приложения указывается information extraction. 16% всех ресурсов из этих 1200 позиционируются как ресурсы для извлечения «знаний». На втором месте — машинный перевод, 12%, но на следующем месте собственно извлечение знаний, то есть это еще 7%, итого 23%. Они разделяют information extraction и knowledge discovery. Терминов много. Это данные программного комитета конференции LREC. С большим отставанием идут «лингвистическое моделирование», «теоретические базы данных» и прочие направления, направленные на доказательство лингвистических теорий, например. То есть это уже по 4—5%, но среди этих мелких задач возникает еще некоторое количество близких нам тематически. Например, то, что называется emotion recognition или sentiment analysis, есть такой термин.

В.Ф. ХОРОШЕВСКИЙ: Sentiment analysis — это анализ тональностей.

В.П. СЕЛЕГЕЙ: Да, то есть тоже в некотором смысле извлечение информации, это еще, между прочим, 4%. Эта тема находится на 6-м месте. 4% ресурсов упоминают такую, казалось бы, специфическую область в качестве важнейших для своего ресурса. Дальше named entity recognition, еще несколько мелких

подразделений, которые сегодня упоминались. Если их просуммировать, получится такая огромная цифра в 38%. То есть 38% ресурсов позиционируются как ресурсы для извлечения знаний или информации. Это, конечно, очень серьезное изменение, раньше такого не было. Синтез речи — 1%, natural language generation — 1%, морфологические парсеры — 2%, классификация документов — 3%, semantic web — 3%, language identification — 3% , хотя это задача, которая тоже вроде к нам относится… То есть это уже 40%.

ВОПРОС ИЗ ЗАЛА: Opinion mining?

В.П. СЕЛЕГЕЙ: Ну это вошло в какую-то группу из упомянутых. Такая ситуация, что 40% всех ресурсов, которые создаются, имеют задачу извлечения знаний как одну из основных.

К.В. АНИСИМОВИЧ (Компания «ABBYY»): Я немного хотел бы пролить свет на первый пункт, на отличия между information extraction и knowledge acquisition. Как, собственно говоря, достаточно подробно рассказала Лена Козеренко, knowledge acquisition — это область, которая берет начало из искусственного интеллекта и посвящена наполнению знаниями базы знаний искусственного интеллекта. При этом существуют как автоматические методы machine learning, так и ручные или комбинированные методы, которые традиционно именуются knowledge engineering. Принципиально то, что пользователями этих знаний является не человек, а машина вывода системы искусственного интеллекта. Это основное различие от information extraction, потому что если мы говорим об information extraction, мы говорим про автоматическое извлечение фактов, данных знаний из текстовых корпусов, которым пользуется человек, как яндексовская служба «Пресс-портретов», то есть это информация, которая извлекается и представляется в human readable форме. А если мы говорим о knowledge acquisition, то это прежде всего автоматизированное ручное или полностью автоматическое обучение системы искусственного интеллекта, а полученные знания будут использоваться не человеком, а машиной вывода. Это вполне традиционная точка зрения, которая во всех учебников по искусственному интеллекту, по-моему, проводится.

А другой вопрос — если knowledge acquisition осуществляется по текстовым корпусам, то практические методы очень похожи на методы, которые использует information extraction.

А. РЫЛОВ (Компания «ABBYY»): У меня два практические вопроса к Елене Игоревне. Вы сказали, что на западе кто-то делает deep approach, соответственно, интересно, кто. А второй вопрос к Дмитрию Владимировичу. Вы

сказали, что одна из проблем — проблема скорости. Какой тогда критерий можно взять, что такое быстро, что такое медленно с точки зрения заказчика. Вот Ontos — это быстро, медленно, и вообще как сравнивать?

Д.В. ЛАНДЭ: Я не могу сразу ответить. Я знаю, что технологически процедура индексирования баз данных проходит на порядки быстрее, чем процедура извлечения данных. Я только это могу констатировать. Сказать, что такое быстро? Наверно, то, что нужно заказчику.

В.Ф. ХОРОШЕВСКИЙ: Это в технологическом цикле, или это вынесено за технологический цикл? Если вы семантически индексируете за технологическим циклом использование информации, это один вопрос. А если вы на лету индексируете, это совсем другой вопрос. Тогда у вас время отклика должно не раздражать пользователя. Что такое не раздражать? 1—3 секунды. Всё. Если он ждет 10 секунд, он ушел.

Д.В. ЛАНДЭ: Ну, 1—5 секунд. Если очень нужна информация, можно и подождать.

В.Ф. ХОРОШЕВСКИЙ: Смотря какая информация. А вот когда нужно обработать корпус большой, чтобы потом быстро отвечать на вопросы, вот это можно и подождать.

Е.И. БОЛЬШАКОВА: Я постараюсь ответить на вопрос. На самом деле я не знаю коммерческих систем, работающих на deep, я могу только судить, какие ведутся исследования. Но очень большое количество исследований в области information extraction, они так позиционируются, идут в соответствующую секцию. У них как бы постановка задачи такая, что у них уже есть разбор. И после этого они начинают спокойно извлекать. Глядя на то, к чему приводит поверхностный, я говорю, может быть, и нам поменять постановку задачи?

В.Ф. ХОРОШЕВСКИЙ: У меня одно возражение. Поскольку нет анализаторов, которые сейчас для полного языка обрабатывают всё, у вас нет другого выхода, как просквозить между вот этими столбами, когда, с одной стороны, вам нужно получить результат, с другой стороны, вы естественно по мере увеличения правил они начинают кусаться. Как только они начинают кусаться, у вас появляются статистические, эвристические оценки, начинаются такие вот идеи, как скрестить ежа с ужом и не получить при этом десять метров колючей проволоки, а что-нибудь более интересное. И вот тогда вы говорите: «Ага, а

давайте я по-быстренькому именную группу выщелкну. И эту именную группу я по полной разберу». Правильно. Вот сейчас, на мой взгляд, почти все уже так и делают. И вопрос-то, для меня по крайней мере, deep — это когда стройными рядами иду — морфология, поверхностный синтаксис, глубинный синтаксис, и так далее по всей цепочке, как нас учили патриархи.

ВОПРОС ИЗ ЗАЛА: А где deep начинается?

В.Ф. ХОРОШЕВСКИЙ: Это вопрос к Игорю Мельчуку.

Е.И. БОЛЬШАКОВА: Так вот на самом деле интересно, а все-таки до какой степени мы должны глубоко обрабатывать тексты, чтобы получить приемлемые результаты, и это самый интересный вопрос, который, я не уверена, что будет решен в ближайшее время.

В.Ф. ХОРОШЕВСКИЙ: Спросите у Яндекса, он вам ответит. Когда пользователь будет это покупать.

Е.И. БОЛЬШАКОВА: Нет. На самом деле относительно существования хороших синтаксических анализаторов, мне кажется, они есть, только они скрываются. Я передаю микрофон лингвисту, который, возможно, сильно дополнит мой ответ.

Н.В. ЛУКАШЕВИЧ: Я, возможно, попробую ответить на вопрос по поводу того, что есть из продуктов, реализующих глубинный подход на Западе. Ну, наверно, я буду говорить про университетские исследовательские проекты. Потому что, что касается коммерческих проектов, это отдельная история. Больше всего ссылок и применений в задачах, связанных с извлечением и обработкой знаний, такого исследовательского проекта, как Ontology Works. Автор проекта — Бонни Дор и ее группа, это, я в общем понимаю, успешно развивающийся проект, и на основе этого подхода много чего сделано… Я в свое время занималась анализом того, какие проекты поддерживает NSF, связанные с задачами, которыми занимаемся мы. И вот проект по Ontology Works тогда получил двухмиллионное бюджетное финансирование в 2004 году. Я дальше его отслеживала. Но на самом деле все, что касается deep approach, связанного с применением каких-то онтологических и глубинно-семантических методов, ссылаются на эти фундаментальные исследования этой группы и, более того, они уже реализовали проекты. То, что я назвала, называется Ontology Works, это некоторый framework…

А.Ю. НЕДОЛУЖКО (Карлов университет, Прага): У меня несколько вопросов. Первая деталь — Владимир Федорович в начале показал схему, что делают все, что не делает никто, я бы хотела вступиться за анафорические ссылки. На самом деле они есть и в MUC'ах, особенно в последних — это одна из пяти задач…

В.Ф. ХОРОШЕВСКИЙ: Когда я говорил про анафорические ссылки, я прежде всего имел в виду российскую действительность. Потому что когда мы говорим об английской анафоре, особенно если это местоименная анафора, там с одушевленностью, за исключением субмарин, как известно, гораздо проще, чем в русском языке. А вот когда вы в русском языке пытаетесь даже местоименную анафору анализировать, у вас гораздо больше проблем. Я уж не говорю о более сложных лингвистических явлениях.

А.Ю. НЕДОЛУЖКО: Я исключительно в мировом масштабе…

ВОПРОС ИЗ ЗАЛА: А насколько успешны они? Вы сами проверяли?

А.Ю. НЕДОЛУЖКО: Я сама это делаю, но не на английском, а на чешском, поэтому я постоянно не читаю. Но я их видела.

ВОПРОС ИЗ ЗАЛА: А какая точность навскидку?

А.Ю. НЕДОЛУЖКО: Около 90, 89, наверно. То есть больше, чем 85.

В.Ф. ХОРОШЕВСКИЙ: По точности или по полноте?

А.Ю. НЕДОЛУЖКО: К общим вещам, к ситуации «Россия versus остальной мир». Если мы чуть-чуть перевернем и поставим не «Россия versus остальной мир», а «английский язык versus остальной мир», то, возможно, Россия — будет не так плохо. Действительно, то, что сделано на английском языке сделано настолько больше, чем на других языках. То есть на других языках так же спорят о стандартах, как и мы спорим о стандартах, и много сделано всего, и посчитано всего, и больше существенных оценок для information extraction, но все равно, возможно, сравнивать русский с английским не совсем корректно.

Еще один вопрос к Лёве как представителю бизнеса. Ты сказал, что вопрос о том, нужна ли анафорическая разметка, вытекает в процессе решения задачи, так вытекает, что нужна или не нужна? Это мне важно знать.

Л. ГЕРШЕНЗОН: Может быть, я очень резко сказал про анафорические связи и их установления, это моя любимая задача, мы с Лёшей Сокирко этим занимались сколько-то месяцев, это очень дорого, если перевести на деньги. И Лёша потом был оппонентом, в общем, защитилась в связи с этим кандидатская диссертация, короче, это очень близкая, родная и любимая мне тема. Лёша получил очень хорошие результаты в какой-то момент. Это было пару лет назад. У нас возникли серьезные проблемы с внедрением этих результатов. Откуда эта задача для информационного поиска, это понятно, есть этот TF-IDF, а есть белые пятна, которые непонятно, что такое, они вообще-то стоп-слова, местоимения, а если их заменить на какие-то значимые слова, то, может быть, у нас качество поиска улучшится. Может быть, это, конечно, правда, но когда мы говорим про злобный бизнес, есть вопрос не только производительности, но и цены. В тот момент оказалось, что эти хорошие результаты сложно внедрить, потому что это довольно сложная система. Оказалось, что есть много чего другого, что эту проблему может решить. Другое дело, что, скажем, для отдельно любимого мной закрытого проекта поиска цитат в новостях — это очень полезная и нужная штука. Есть база цитат, кто говорит, что говорит. Все знают, что конструкций типа «он сказал», «по его мнению» в текстах очень много, мы считали, до 30%, если по всем цитатам смотреть. Это реально большой, очень ценный кусок полноты. В этом месте это действительно очень полезная и необходимая штука. У нас очень много чего хочется сделать, не на всё хватает рук. Поэтому вопрос, нужно — не нужно, для каждой конкретной задачи решается по-своему. В поиске пока можно жить без этого, на наш взгляд, а завтра уже по-другому, вот и всё.

Н.В. ПЕРЦОВ: Вопрос к Льву Гершензону или кому-то из присутствующих. Идентификация цитат в тексте — это information extraction или knowledge acquisition? Это чужие слова или извлеченные цитаты из какого-то известного источника — это что?

Л. ГЕРШЕНЗОН: Я должен признать, что я пытался перед круглым столом еще у некоторых коллег выяснить эту разницу…

Б.В. ДОБРОВ (НИВЦ МГУ им. М.В. Ломоносова): Я начну с дискуссионного вопроса. А что мешает собравшимся организовать нечто вроде РОМИПа или дорожку в рамках РОМИПа для того, чтобы начать решать поднимаемые вопросы? Теперь более конкретно. Действительно справедливо говорилось, что

движение a la TREC начиналось еще раньше и оно началось по заказу. Однако есть европейская конференция и азиатская, и РОМИП сам родился в гораздо меньшей степени по заказу… Даже существовал без Яндекса, на средства участников. Потому что смысл участия в РОМИПе заключается в том, что с каждым годом результаты участников улучшаются. Кроме того, что мы видим сейчас, когда обсуждаются темы information extraction и knowledge acquisition? Что реально много не устоялось, задачи четко не поставлены, как справедливо здесь отмечалось во многих выступлениях — ну сделали модуль information extraction, ну и что? На самом деле он сам по себе не работает, и никто его не покупает, потому что он реально является частью других систем. А каких систем? На самом деле, существует огромное количество разных вопросов. Реально соревнования типа РОМИП или TREC приводят к тому, что терминология уточняется, постановки задач уточняются. На самом деле в РОМИПе мне больше всего нравится, что люди публикуют не там какие-то результаты лучше всех с мутным объяснением, что они сделали, а честно публикуют плохие результаты и четко объясняют, почему они так получили. Это экономит остальным участникам значительные деньги. Поэтому совсем не обязательно нужен внешний заказчик, а польза из этого достигается. Что нам мешает устроить РОМИП?

В.Ф. ХОРОШЕВСКИЙ: Я попытаюсь ответить. На мой взгляд, мешает ровно две вещи. Первая — это боязнь того, что конкурент перехватит результат и быстрее получит за это деньги. Результат — это алгоритмы, которые использованы для того, чтобы получить качество. В том же РОМИПе и в том же TREC’е есть два типа. Человек может публиковать алгоритм или только результаты. Если он публикует результаты и объясняет, почему у него получились хорошие или плохие результаты, из этого умный конкурент извлечет алгоритм.

Б.В. ДОБРОВ: Ситуация заключается в том, что, если вы люди рынка, то ваша позиция стоит не на том, что вы реализовали какой-то хитрый алгоритм на коленке, а в том, что за вами стоит десятилетие человека труда.

В.Ф. ХОРОШЕВСКИЙ: И это десятилетие стоит денег.

И.В. СЕГАЛОВИЧ (Компания «Яндекс»): Ну как бы его не перехватишь просто так. Хорошо, вы можете перехватить у «Яндекса» Matrix.Net? Понадобится миллиард документов и огромная размеченная коллекция.

В.Ф. ХОРОШЕВСКИЙ: И второе, на мой взгляд, не менее важное. Это то, что отсутствуют по-настоящему evaluation метрики. Если говорить о метриках TREC’a,

MUC'a, DUC'a, то это в настоящий момент уже детский лепет на лужайке. Потому что те метрики, которые там используются, они не отражают действительности ни в коей мере, а новых метрик практически нет.

Б.В. ДОБРОВ: Когда задается вопрос, что метрики плохие, явно подразумевается, что TREC или РОМИП являются соревнованием. Относиться к этому так совершенно нельзя. Это бессмысленно. Хорошо, мы участвуем в РОМИПе, ну что, по-видимому, мы соревнуемся с «Яндексом», с компанией в 300 миллионов долларов или сколько у них, полмиллиарда долларов оборотом? Нет, конечно. Речь идет о совсем другом. Метрики — это лишь инструменты померить некие задачи в искусственной обстановке. Они предназначены для того, чтобы уточнить постановки задачи и сэкономить усилия. Не более того. Это не соревнование. И иначе зачем мы это сегодня говорим? Если нет ни метрик, ни постановок задачи…

В.Ф. ХОРОШЕВСКИЙ: Была статья, специально посвященная метрикам, но никто на эту статью не обратил внимания. Ну это к слову. А почему плохие метрики? А потому что вы меряете объект, но не меряете атрибуты объекта и многое другое. А уж если вы говорите о фактах, но не меряете отношений и не меряете ни точности, ни полноты отношений. Тогда о чем мы говорим? Вы спрашиваете: «Давайте посоревнуемся в TREC'е»… Не посоревнуемся, сравнимся, обсудим извлечение трех типов сущностей — люди, организации, location'ы.

Б.В. ДОБРОВ: В РОМИПе, кто-то, может быть, и соревнуется, но реально правильное слово — тестируют.

В.П. СЕЛЕГЕЙ: Но вы заинтересованы на самом деле в такого рода тестировании? Вы говорите, что метрики никуда не годятся, но вас это разочаровывает как разработчика?

В.Ф. ХОРОШЕВСКИЙ: Как разработчика меня сильно разочаровывает. Мне на самом деле идея Ермакова, который предложил на РОМИПе, насколько я помню, дорожку для information extraction, и никто его не поддержал, и поэтому, собственно, один участник — неинтересно. Меня это очень пугает, убивает и всё остальное. Потому что в действительности чтобы идти дальше, нужно видеть не только свой садик-огородик, а то, что происходит у других. И с этой точки зрения мне было очень важно бы участвовать в такого типа тестировании. Но обращаю ваше внимание, что в любом случае это все равно соревнование — ума, работы, вложений и всего остального. И мне неинтересно тестироваться или

соревноваться, если метрики не отражают моего богатства. А получается, что те метрики, которые меряются, они, в общем-то, рассчитаны на гораздо более низкий уровень, чем тот, который достигнут был… но я, к сожалению, не могу выступать полномасштабно от имени компании «Avicomp», но те вещи, которые сделаны в «Avicomp'е», они, вообще говоря, мощнее, чем то, о чем здесь говорилось. И по information extraction, и по идентификации.

И.В. СЕГАЛОВИЧ: На самом деле метрики, которые у нас используются в РОМИПе, они обсуждаются участниками и меняются. Я помню, года два назад очень критиковал какую-то метрику, и все договорились, что у нас появится еще одна. Никто не мешает новому участнику на форуме в свободной дискуссии предложить другую метрику, ту, которая ему больше нравится. Просто некоторое непонимание в этом месте. И второй момент — хочется уточнить насчет количества участников. Все кто может, все, кто хочет себя измерить-потестировать, там нет никакого давления, нет никакого, как говорит Борис, соревнования, чемпионата. Пожалуйста, можете несколько прогонов отдать — тестовый, базовый, продвинутый, слабый, примитивный. Посмотреть, как они работают, и договориться о метриках. То есть ценность основная и РОМИПа и, как я подозреваю, старых конференций со старыми метриками — в том, что появляются корпуса. Я обращаю ваше внимание на то, что TREC и особенно РОМИП — это самый большой корпус с оценками asessor'ов по запросам и по фактам. Это бесценная вещь. Мы все жалуемся, что у нас не хватает хороших команд, хороших исследований — у нас их очень сильно не хватает, во многом потому что нет больших корпусов, в которые вкладывались даже данные про те же западные государственные организации американские военные. Они понимают, что большой корпус — это тот же речевой корпус. У нас вообще ноль, по большому счету, в России размеченного речевого корпуса. Потому что не вкладывается, деньги государственные уходят в песок. Вот РОМИП — это правильная модель, результатом РОМИПа является бесценная вещь: можно взять эти корпуса и натренировать свою систему, за бесплатно. За ноль рублей ноль копеек вы получаете общественное достояние. Поэтому я призываю всех специалистов, работающих над information extraction и knowledge acquisition, присоединиться к инициативе РОМИПа и поиграть… по сути это детская игра, детская песочница, но на выходе возникает серьезный материал для всех, общее достояние. А соревнование — у кого лучше, у кого хуже — этого нет, в РОМИПе в уставе запрещено называть, что «я победил» и пиарить свою победу.

В.Ф. ХОРОШЕВСКИЙ: Но в TREC'е, MUC'е и DUC'ах это эксплуатируется нещадно. Если вы попали в пятерку лучших в TREC'е, это означает, что контракты достаточно мощные будут ваши.

И.В. СЕГАЛОВИЧ: У нас с самого начала были Rambler, Mail… даже Mail’а не было, Rambler, Aport и Яндекс, и было очень мало научных групп, потому что научные группы приносили статьи на «Диалог», которые доказывали, что их поиск — самый лучший поиск в мире, а в РОМИПе они не участвовали, но как бы вот эти три игрока договорились, что мы соревнуемся у водопоя, не будем грызть друг другу горло, а будем спокойно тестировать, как Борис говорит, свои системы, и мы договорились с самого начала, что мы не будем использовать для пиара, иначе это превратится в черт-те что. И к счастью это уже семь лет соблюдается.

В.Ю. АНТОНОВ (МГУ им. М.В. Ломоносова): Я немножко другое хотел сказать. Вот было сказано, что заказчиков нет, и поэтому соревнований нет. На самом деле заказчики есть, потому что хотя бы вспомните уже упоминавшийся здесь «Кронос». Какая база у него здесь в России — огромная совершенно база. Естественно, что эти базы получены совсем не лингвистическими, не научными методами... Но эти базы надо пополнять. Пополнять, верифицировать — еще одна задача возникает.

В.Ф. ХОРОШЕВСКИЙ: Этим занимается один из проектов «Avicomp’а».

В.Ю. АНТОНОВ: … Заказчики есть. Ну и хотелось сказать, что на самом деле сказано о том, что всё хорошо, в частности извлечение такими простыми методами объектов, конечно, здесь очень большая проблема как раз отождествления и разделения объектов. Потому что все эти хорошие цифры, которые здесь назывались, они без учета этих вот вещей.

Л. ГЕРШЕНЗОН: Можно я все-таки отвечу, что есть заказчики? Селегей говорил, что задачи sentimental analysis, opinion mining — сколько там, 4%? Очень много в сумме. Откуда так много? Больше в четыре раза, чем генерация речи. А потому что есть такая штука в Америке, как customer reporting, и большие компании не могут без этого обойтись. Они должны огромные объемы этого дела анализировать. И от этого зависит их бизнес. Хоть одну такую организацию в России можете назвать, которой важно автоматизированно обрабатывать отзывы пользователей, у которой от этого что-то зависит?

РЕПЛИКА ИЗ ЗАЛА: «МТС».

РЕПЛИКА ИЗ ЗАЛА: Те, кто торгуют на бирже, они используют.

Л. ГЕРШЕНЗОН: Отзывы пользователей. То есть я хочу сказать, что сервисные компании — ни «Билайн», ни «МТС»… Я хочу сказать, что в этой целой области 4% от лингвистических ресурсов. Ну, может быть, игрушка. «Ой, здорово, у нас будут отзывы, или у нас их не будет». Есть — хорошо, нету — и так проживем. А развитие этого, пока это тут кому-то не будет жизненно необходимо, на мой взгляд, не будет.

А.С. СТАРОСТИН (МГУ им. М.В. Ломоносова): Я с тобой согласен, но я хочу причину указать. Мне кажется, что во многих больших компаниях, которые в России уже возникли, достаточно низкий уровень, ну я не знаю, бизнес-культуры, что ли, или еще чего-то, очень многие еще действительно не понимают, что они могли бы перерабатывать большие коллекции текстов, и это бы действительно им принесло богатство. Они действительно до этого не доросли.

В.Ф. ХОРОШЕВСКИЙ: Когда от этого зависит, останется он на второй срок или нет, они очень хорошо понимают.

?: Вообще говоря, задача такая есть, задача очень популярная, задача называется анализ информационного риска. Другой вопрос, что решать эту задачу проще руками сегодня. С помощью Яндекса. А дальше ручками. И решают, и решают много.

С. ПРОТАСОВ (Компания «Рамблер»): Хотел сказать немножко про метрики. Да, действительно, у нас нет хороших метрик. Одну из них мог бы предложить — это реально работающий сервис, на который ходят много пользователей. Например, «Покупки»…, либо «Рамблер-вакансии». Это довольно хороший критерий, с моей точки зрения.

РЕПЛИКА ИЗ ЗАЛА: Посещаемость — это следствие.

В.Ф. ХОРОШЕВСКИЙ: Опосредованно — да. Как связать?

С. ПРОТАСОВ: Ну вот если вы умеете извлекать знания и факты из Интернета, то вы сможете для себя это проиндексировать и потом использовать.

В.Ф. ХОРОШЕВСКИЙ: А если вы публикуете порнушку, то, ничего не извлекая, вы получаете пользователей гораздо больше.

С. ПРОТАСОВ: Я рассказываю про конкретный ресурс, который дает пользователям структурированную информацию.

Н.В. ПЕРЦОВ: Я хочу задать вопрос. Основным противопоставлением, которое здесь звучало, между извлечением знаний и тем, что я не знаю, как хорошо передать по-русски, knowledge acquisition. Вот для information extraction существует ходовой эквивалент — извлечение знаний, и это входило в название круглого стола. А для knowledge acquisition существует или нет?

РЕПЛИКИ ИЗ ЗАЛА: Приобретение знаний.

Н.В. ПЕРЦОВ: Проблема связана с терминологией. Может быть, об этом можно будет поговорить завтра, послезавтра или в кулуарах, но я хочу сказать, что с терминологией, а стало быть, с осмыслением обозначаемых сущностей дело обстоит неблагополучно. Я задал вопрос представителю бизнеса, является ли то, о чем он говорил, а именно фиксация некоторых фрагментов текста как цитат или передача некоторой чужой речи, является ли это приобретением знаний или извлечением знаний. Он не смог ответить, и никто пока мне на этот вопрос не ответил.

Е.И. БОЛЬШАКОВА: Это скорее всего извлечение информации.

Н.В. ПЕРЦОВ: Тогда верно ли, что то, что получается в результате логического вывода из текста, является приобретением информации?

Е.И. БОЛЬШАКОВА: Ну наверно.

Н.В. ПЕРЦОВ: Наверное! Вот в чем дело! Дело все в том, что на самом деле у знатоков в этой области нет четкого представления, что является чем. Я хочу сказать, что я готовил выступление, я приведу только два эпиграфа к этому выступлению — первый, это эпиграф к книжке Мельчука «Курс общей

морфологии», который заканчивается: «Ах, доктор, так ли уж важна терминология?» По-английски это звучит: «Is terminology that important?»… «В терминах ли дело?» И второе — один из присутствующих поймет, что я имею в виду, — «Зачем употреблять неуклюжее слово “последовательность”, когда есть хорошее слово “кортеж”?» Это очень хорошая цитата, которая говорит, что в русский язык можно привлекать слова, транслитерированные из других языков. Я считаю, что в русский язык очень хорошо ввести слово mining. Потому что data mining — это совсем не то, что information extraction. А то начинается facts mining и так далее, и получается какой-то невероятный конгломерат терминов, причем непонятно иногда, что имеется в виду. Русский язык очень общежительный, он охотно принимает транслитерированные слова, давайте говорить тогда по-русски — «майнинг», а не «mining».

Л. ГЕРШЕНЗОН: Я думаю, это можно считать итогом круглого стола, что давайте больше обращать внимание на русских язык.

В.Ф. ХОРОШЕВСКИЙ: Что в действительности произошло и что, на мой взгляд, достаточно интересно? Это то, что мы попытались в меру своих знаний и понимания ответить на некоторые вопросы. Были противоположные мнения, и они высвечивали некоторые аспекты той проблемы, которую мы обсуждали. На мой взгляд, очень важно по сухим остаткам от сегодняшнего обсуждения было бы действительно образование такой дорожки по information extraction, но это потребует от ее организаторов огромных усилий и по получению стандарта, и по разработке системы оценки. И второе, что я хотел бы отметить, что, к сожалению, суровая проза рынка убивает те идеи, которые были изначально заложены и в «Диалог», и в научное содружество людей, которые были рождены в Советском Союзе. И сейчас это очень сильно, на мой взгляд, влияет на развитие науки. Как это преодолеть, надо вместе пытаться это сделать. И если не мы, то кто? Спасибо всем.