XXIV Европейская летняя школа по логике, лингвистике и информатике (ESSLLI 2012)

6–17 августа 2012 г. в Ополе (Польша) прошла XXIV Европейская летняя школа по логике, лингвистике и информатике (ESSLLI 2012). Ежегодно, начиная с 1989 г., это мероприятие проходит под эгидой Ассоциации логики, лингвистики и информатики (FoLLI) в одном из университетов Европы. В нынешнем году вниманию участников школы было предложено 48 лекционных курсов и семинаров, по традиции подразделенных на три «интерфейса»: «Language and logic», «Language and computation» и «Logic and computation». Наш обзор, нацеленный на читателя-лингвиста, рассказывает о ряде курсов из первых двух тематических блоков, т.е. дисциплинарно относящихся к формальной семантике и к компьютерной лингвистике.

Курсы интерфейса «Language and logic» объединяет ориентация на анализ естественного языка теми или иными логическими средствами, однако мотивация исследователей неодинакова: в одних случаях анализ нацелен преимущественно на непротиворечивое и максимально полное семантическое описание, в других во внимание принимаются и (аналитико-)философские проблемы, такие как онтологический статус множеств и интерпретация модальностей.

Курс «Multidimensional semantics», прочитанный Л. Превo и Л. Вьё (Франция), имел целью представить различные подходы к анализу естественного языка, основанные на выделении более чем одного «измерения» в семантическом представлении предложения. В число явлений, анализируемых таким образом, входят: личные местоимения и дейктические слова типа *сегодня*, *здесь*; прямая, косвенная и несобственно-прямая речь (direct, indirect, mixed quotations); пресуппозиции; вставные конструкции и сообщающие фоновую информацию клаузы (supplements); структура дискурса в рамках различных версий теории репрезентации дискурса (Discourse representation theory, DRT).

Понятие «измерения», ключевое для курса, не имеет единственного строгого определения и варьируется от подхода к подходу. Прототипическими случаями «многомерности» в семантике могут считаться те, когда интерпретация выражения зависит от нескольких эксплицитно указанных индексов (к примеру, возможного мира и временно́й точки соотнесения в грамматике Р. Монтегю), однако расширение инвентаря исходных семантических типов (скажем, за счет типа «utterances» у К. Поттса) также ведет к «многомерности». Нет единства и в отношении того, является ли взаимонезависимость двух компонентов значения необходимым условием выделения их в разные «измерения». Открыт также вопрос о соотношении понятий «измерение» (dimension) и «уровень» (level). Так, дистинкцию Д. Каплана для местоимений «character / content» естественнее понимать как отражение различных уровней или этапов в процессе композиции значения (вначале character вместе с индексом контекста определяет content, который вместе с индексом возможного мира дает экстенсионал), соотношение пропозиционального содержания высказывания и его роли в дискурсе – как низший и высший уровни анализа (синтаксис vs. текст). В то же время «double contribution» (явление, при котором цитируемый фрагмент участвует в семантике предложения одновременно как звуковая цепочка и как значимая единица речи, ср. *«My girlfriend bought me this tie,» said John, but I don't think she did* у Б. Парти) или сочетание в одном высказывании нескольких дискурсивных функций, напротив, представляют собой более или менее независимые «измерения».

Основной проблемой, затронутой в ходе курса «Plurals in semantics and philosophical logic» С. Флорио (США, Великобритания) и О. Линнебо (Великобритания), был поиск корректного семантического представления для именных групп с семантическим признаком ‘множество’. Среди трудностей, с которыми сталкивается всякий анализ, недистрибутивность (ср. англ. *Our group met*, *Tom and John met*, но **Tom met*) и некумулятивность (из *Эти линии параллельны* и *Те линии параллельны* не следует *Эти и те линии параллельны*) некоторых предикатов, а также логико-математические парадоксы теории множеств (ср. пример М. Резника *There are some sets that are self-identical, and every set that is not a member of itself is one of them*, представляющий собой истинное высказывание, но приводящий к парадоксу Рассела, если понимать *some sets* как обозначение некоторого множества множеств).

Авторы курса предлагают в качестве решения названных проблем так называемую «логику множественности» (plural logic), в которой вводятся особая синтаксическая категория множественных термов с множественной референцией и квантификация по ним. Таким образом, в примере Резника *some sets* имеет множественную референцию к множествам, а не единичную референцию к «множеству множеств». Один из вариантов plural logic располагает также предикатами, способными

принимать специфически множественные аргументы, что позволяет отразить недистрибутивность. Так, из *Холмс и Ватсон знакомы* можно вывести $\exists xx$ (знакомы *xx*), где *xx* – переменная, пробегающая по множественностям, а предикат знакомы выполняется только множественностями (**Холмс знаком*).

Ставится также вопрос о «надмножественных» (superplural) термах и квантификации, причем аргументами в их пользу считается допустимость предложений типа ***Эти люди и те люди*** *соревнуются друг с другом* с предполагаемой интерпретацией «соревнуются две команды (т. е. множественность, составленная из двух множественностей), а не их участники один на один», а также существование в естественных языках единиц типа лит. *dveji* 'две [пары]'. Примеры такого рода, так же как и постулируемое преимущество plural logic перед обычным теоретико-множественным подходом в отношении парадоксов, вызвали дискуссию, не принесшую окончательного ответа.

Курс «The semantics of attitude reports» Э. М а й е р а (Нидерланды) был посвящен проблемам высказываний о пропозициональных установках (propositional attitude reports) – высказываний, имеющих форму *X полагает / надеется / желает / ..., что(бы) p*, где *X* – обозначение субъекта установки, а *p* – выражение некоторой пропозиции. По этой теме существует огромное количество литературы, причем первым из авторов, уделивших ей внимание, следует, по-видимому, считать Г. Фреге, использовавшего дистинкцию смысла (Sinn) и значения (Bedeutung) для объяснения того факта, что, хотя *Утренняя звезда* и *Вечерняя звезда* обозначают один и тот же объект, (1) может быть истинным при ложности (2):

1. *Иван полагает, что Утренняя звезда – планета.*

2. *Иван полагает, что Вечерняя звезда – планета.*

В дальнейшем Б. Рассел различил у предложений типа *Георг IV желал знать, являлся ли Скотт автором «Веверлея»* чтения *de re* (вопрос о конкретном человеке, которого Георг IV в своем вопросе мог даже не назвать Скоттом, а, скажем, попросту указать на него) и *de dicto* (вопрос о том, являются ли тот, кого называют Скоттом, и автор «Веверлея» одним и тем же лицом). Наряду с этими двумя чтениями, существуют пропозициональные установки *de se*, т.е. приписывание субъектом тех или иных свойств самому себе; в таких случаях важно отразить в формализации отождествление субъектом установки носителя приписываемого свойства с самим собой.

На занятиях курса обсуждались три подхода к проблеме *de dicto / de re / de se*. Один из них, синтаксический, связывает чтение *de re* с происходящим на уровне логической формы особым передвижением составляющей (*res* movement), позволяющим ей получать более широкую сферу действия (scope), чем у интенсионального глагола. В варианте Д. Льюиса, постулирующего дополнительно отношение знакомства (acquaintance) между субъектом и объектом пропозициональной установки, этот подход способен моделировать и установки *de se*: требуется лишь уточнить отношение знакомства как тождество. К другой группе относится подход Д. Каплана к *de re* и *de se*, основанный на представлении местоимения первого лица в его двумерной семантике, в которой интерпретация выражения зависит от контекста употребления, включая говорящего, и от возможного мира (см. выше). Третий подход – основанная на идеях Х. Кампа теория репрезентации дискурса, удобно представляющая различия в сфере действия, отличающие *de dicto* от *de re*, и обладающая несколько иным, нежели в стандартной логике предикатов, механизмом связывания (binding). На стороне теории репрезентации дискурса, вероятно, находятся симпатии автора курса: ей было уделено больше всего времени, и автору удалось показать совместимость DRT с некоторыми идеями Каплана и представимость в ней отношения знакомства в качестве вводимой пропозициональной установкой пресуппозиции, откуда следует, что DRT может моделировать все три типа установок: *de dicto*, *de re* и *de se*.

В завершение курса были затронуты и глаголы пропозициональной установки, отличные от эпистемических: *hope* 'надеяться', *desire* 'желать' и др. В предложенной модели эпистемическая установка может выступать основой для прочих (так называемые установки *de credito*).

В компьютерно-лингвистической части школы значительная часть курсов была так или иначе ориентирована на корпусные исследования. Р. Ш е ф е р и Ф. Б и л ь д х а у э р (Германия) прочитали вводный курс «Building large corpora from the web» о методах сбора текстовых коллекций объемом порядка нескольких миллиардов словоупотреблений, их автоматическом аннотировании и применении в задачах лингвистики. По мысли авторов курса, веб-корпус одного языка должен быть основан на случайной выборке из электронных документов, написанных на этом языке и открыто доступных в сети. Чтобы добиться почти случайного отбора страниц, удобно настроить поисковый робот (crawler) на обход веб-документов по гиперссылкам, начиная с нескольких страниц, выбранных как стартовые (seeds). Весь материал, полученный таким способом, будет первым

делом нуждаться в технической постобработке: приведении к единой кодировке, очистке от метатекстовой разметки, удалении документов на посторонних языках, исключении слишком коротких текстов, дублирующихся фрагментов и др. Эти задачи требуют большой вычислительной мощности, но для их решения уже известны эффективные приближенные алгоритмы.

Когда технический этап завершен, предстоит лингвистическая постобработка. Здесь первая проблема – токенизация: ни один существующий токенизатор не является достаточно общим, чтобы правильно обрабатывать любые веб-документы, из-за чего в корпусе появляется большое число гапаксов. В результате страдает качество лемматизации и (статистического) морфологического анализа. Дополнительную трудность лемматизатору создают орфографические ошибки и опечатки; авторы курса предлагают обходить эту помеху добавлением нового уровня разметки – «орфографически нормализованного» текста. Анализ готового веб-корпуса (в частности, анализ распределений словоформ и предложений по длине) позволяет устранить просчеты в стратегии обхода страниц и в постобработке, а чтобы содержательно сравнить построенный корпус с аналогами, хотя бы по характеру охваченного речевого материала, оказывается достаточным иметь частотные списки лемм.

Р. Шефер и Ф. Бильдхауэр прежде всего увязывают применение больших корпусов в лингвистике с лингвистическими «редкими событиями» – языковыми фактами, численность которых в корпусах объемом до нескольких сотен миллионов словоупотреблений очень мала (клитизация неопределенного артикля в немецком языке и др.). Авторы курса также показывают, что из достаточно большого случайно сформированного корпуса можно выбрать любой специализированный подкорпус.

Э. Коупстейк (Великобритания) и О. Эрбло (Германия) познакомили слушателей курса «Distributional semantics for linguists» с современным состоянием дистрибутивной семантики. Еще З. Харрис формулирует известную «дистрибутивную гипотезу»: если два слова встречаются в однотипных контекстах, то они близки по значению, – но заметный прогресс в этой области был достигнут лишь около 20 лет назад, когда появилась достаточно мощная вычислительная техника.

Чтобы построить дистрибутивно-семантическую модель, нужен большой лемматизованный корпус. В простейшем случае значение какого-либо слова *w* отождествляется с вектором частот начальных форм, извлеченным из контекстных окрестностей фиксированной ширины (например, по пять словоформ влево и вправо) около всех вхождений форм слова *w* в корпус. Если корпус оснащен синтаксической разметкой, то дистрибутивное представление можно построить несколько сложнее. В обоих случаях нетрудно добиться, чтобы все векторы частот были заданы в одном и том же базисе, и затем снизить размерность полученного линейного пространства (например, удалением низкочастотных элементов базиса). Таким образом строится модель, в которой семантическая близость определена как легко вычислимая мера на векторах: косинус, коэффициент Жаккара и т.п.

Авторы курса отмечают, что понятие семантической близости не соотносится прямо ни с одним из традиционно выделяемых типов лексико-семантических отношений (синонимия, антонимия, гипо-гиперонимия…). Более того, разные ЛСВ одной лексемы и даже омонимичные лексемы имеют единое дистрибутивное представление. Но известен ряд специальных приемов, позволяющих по дистрибутивным данным разграничивать ЛСВ и устанавливать тип семантического отношения между словами.

Одна из лекций была посвящена подходам к объединению дистрибутивной и композиционной (compositional) семантики. Это тенденция последних четырех-пяти лет, однако идеи Б. Кёкке и М. Барони уже приобрели большую популярность среди компьютерных лингвистов. Авторы курса предлагают и свой собственный подход; ключевым для него является понятие «идеальной дистрибуции». Если мир – модель некоторой логической системы, то каждой элементарной формуле, которая истинна в этой системе, можно поставить в соответствие предложение на естественном языке. Дистрибутивно-семантическая модель, вычисленная на основе корпуса из всех таких предложений, и называется идеальной. Непосредственно построить идеальную дистрибуцию можно только для «игрушечных» миров, описываемых очень маленькими онтологиями, но вообще этот путь представляется продуктивным, особенно для анализа квантифицированных высказываний.

Ряд курсов был специально посвящен применению методов машинного обучения (machine learning) к задачам лингвистики. Так, Т. Кисс (Германия) построил свой курс «Annotation mining in R» вокруг одного яркого примера. По правилу немецкой грамматики, существительные в именных группах с предлогом *ohne* ‘без’ должны иметь при себе артикль. Корпусный материал, однако, показывает, что это правило соблюдается далеко не всегда. Какие же причины побуждают носителей языка в одном случае употреблять артикль, а в другом – обходиться без него? Т. Кисс предлагает такое решение:

обогатить разметку контекстов, найденных в корпусе, набором дополнительных признаков (собственная семантика предлога; начальная форма существительного, его словообразовательные, морфологические и семантические характеристики; особенности конструкции всей именной группы и др.) и на полученной базе данных обучить классификатор, который бы явным образом присваивал каждому признаку «вес» – количественную меру значимости. Этим способом удается узнать, например, что каузальная семантика предлога *ohne* и одно или несколько зависимых прилагательных в именной группе повышают шансы употребления артикля, а принадлежность существительного к семантическому классу «отношение» или «признак» понижает шансы. Описанный подход легко обобщить на широкий круг задач не только грамматики, но и лексикологии (таких, например, как снятие неоднозначности). Стоит добавить, что курс Т. Кисса был практико-ориентированным: слушатели учились строить классификаторы средствами статистического пакета R.

Курс А. Сёгора (Дания) «Semi-supervised learning for natural language processing» был весьма техническим по содержанию. В отличие от более известных вариантов машинного обучения «с учителем» (supervised) и «без учителя» (unsupervised), «частичное» обучение предполагает, что на вход алгоритма подается сначала небольшой массив размеченных (labeled) данных, а затем большой массив неразмеченных, при обработке которых классификатор запоминает свои решения и непрерывно пополняет ими тренировочное множество.

Методы, о которых было рассказано в курсе, находят свое важнейшее применение в системах зависимостного синтаксического анализа. В последние годы популярны два статистических алгоритма: (1) нахождение минимального остовного дерева (minimum spanning tree) из взвешенного графа возможных зависимостей и (2) shift-reduce анализ, при котором словоформы из предложения берутся по одной и на каждом шагу выбирается лучшая точка прикрепления начала связи, входящей в данное слово. В первом случае классификатор оценивает веса ребер, а во втором случае – вероятности их прикрепления.

Много внимания автор курса уделил задаче межъязыковой адаптации корпусных ресурсов. Если для какого-либо языка имеется достаточно большой банк синтаксических структур с разметкой зависимостей (dependency treebank), то классификатор, частично обученный на нем с привлечением только морфологической информации, т. е. без учета самих словоформ и их начальных форм, удастся применить для синтаксического анализа близкородственного или даже неродственного, но типологически сходного языка. Качество разбора при этом падает, но не катастрофически.

Ф. Цимиано и К. Унгер (Германия) в курсе «Ontology-based interpretation of natural language» продемонстрировали одну возможность научить компьютер «понимать» небольшой фрагмент естественного языка чисто символическим путем, без использования статистических методов. Авторы курса считают достаточным подготовить два основных компонента: онтологию, представляющую знания человека о какой-либо предметной области, и лексикон, обогащенный синтаксической информацией (валентность глагола и т.п.) в той мере, чтобы компьютер мог без участия лингвиста построить лексикализованную грамматику. Инженер-разработчик должен вручную соотнести каждое понятие, формализованное в онтологии, с некоторой лексической единицей или словосочетанием. В системе Pythia, представленной на занятиях курса, при анализе пользовательского ввода промежуточное представление (между синтаксисом и собственно лексиконом) обеспечивается одним из вариантов формализма DRT. К сожалению, Pythia умеет интерпретировать весьма ограниченный набор высказываний на естественном языке, даже внутри фиксированной предметной области. Тому много причин: неполнота онтологии, сравнительная узость лексикона, обработка только композиционных высказываний, притом корректно построенных в грамматическом и семантическом отношении, и др.

Курс «Empirical approaches to discourse», прочитанный Дж. Спенейдер (Нидерланды), был одним из самых познавательных на летней школе. В центре внимания были когерентные отношения в дискурсе, отвечающие за связанность текста и элементов его структуры, например предложений. Наибольшее внимание уделялось каузальным и контрастивным отношениям. Введением в эмпирические исследования по когерентным отношениям послужили две классические работы по данной тематике (теории Дж. Гоббса, Б. Грош и К. Сиднер). Среди других теорий когерентности в курсе были представлены Rhetorical structure theory (В. Манн и С. Томпсон), разработанная для генерации текстов, и Segmented discourse representation theory (Н. Эшер и А. Ласкаридес).

Дж. Спенейдер указывает проблемы, с которыми можно столкнуться при применении данных теорий когерентности. Например, в основе инвентаря когерентных отношений могут отсутствовать принципы; разные типы информации могут быть объединены в одну

структурную единицу; система аннотации достаточно сложна и пр. Также был дан обзор инициатив по созданию корпусов с разметкой когерентных отношений, в том числе был рассмотрен Penn discourse treebank. В нем выделяется около 100 типов различных отношений, которые могут быть выражены и эксплицитно, и имплицитно. Один из недостатков проекта состоит в том, что в нем не поддерживаются вложенные структуры.

Курс «Natural language processing for historical texts», прочитанный М. Пиотровским (Швейцария), был посвящен автоматической обработке естественного языка в исторических (древних) текстах. За пять лекций, читавшихся в течение недели, был подробно описан процесс «от рукописи к электронному тексту». В основном курс носил технический характер. Слушатели познакомились с различными этапами обработки рукописей и бумажных носителей: сканирование, распознавание текста, конвертирование в различные форматы, дальнейшая проверка текста, программное обеспечение для обработки исторических текстов.

Главный объект курса – исторический текст – понимается как старый текст, значительно отличающийся от сегодняшних и тем самым создающий помехи при его автоматической обработке. От современного текста исторический может отличаться по многим параметрам: вид носителя (пергамент, мрамор, глиняные таблички и пр.), система записи (устаревший шрифт, аббревиатуры и пр.), язык, орфография, также текст может иметь дефекты, быть неясным и т.д. Все эти параметры вызывают определенные вопросы при их передаче в электронную форму: например, необходимо ли сохранять лигатуры или различные типы шрифтов в одном тексте? В ходе курса были подробно обозначены и описаны трудности, с которыми сталкиваются многие проекты по оцифровыванию бумажных носителей.

Много внимания М. Пиотровский уделил оптическому распознаванию символов, в том числе был сделан обзор работ, сравнивающих результаты систем распознавания для исторических текстов. В случае, если текст рукописный и распознать его не представляется возможным, на помощь приходят специальные программы для упрощения работы людей, перепечатывающих рукописи. Их достоинства и недостатки также были рассмотрены в течение курса.

Следующий этап работы с текстом – это его аннотация в соответствии с рекомендациями TEI (Text encoding initiative), исправление ошибок, а также обеспечение многоуровневой разметкой. В последний этап входят такие задачи, как частеречная разметка, лемматизация, парсинг.

В скором времени выйдет в свет книга М. Пиотровского, озаглавленная так же, как и прочитанный им курс лекций.

Д.Б. Тискин, А.С. Шиморина, В.В. Порицкий

Сведения об авторах:

Даниил Борисович Тискин
Санкт-Петербургский государственный университет
tyskin@yandex.ru

Анастасия Сергеевна Шиморина
Институт лингвистических исследований РАН
shinas@yandex.ru

Владислав Валерьевич Порицкий
Белорусский государственный университет
v.poritski@gmail.com