Шестая международная конференция по языковым ресурсам и их оценке (Language Resources and Evaluation Conference – LREC)

Шестая международная конференция по языковым ресурсам и их оценке прошла с 26 мая по 1 июня 2008 года в г. Марракеше (Марокко). LREC – самая большая и разнообразная по направлениям исследований корпусная конференция в Европе. В этом году было принято 645 докладов и участвовало более 1000 человек из стран Европы и Азии, из Канады, США, Австралии и Южной Африки. На официальном сайте www.lrec-conf.org/lrec2008 будут опубликованы все материалы нынешней конференции – организаторы рассматривают это как шаг к развитию мирового корпусного сообщества.

Подобно подмосковной международной конференции «Диалог», которая проводится ежегодно в июне и собирает лучших лингвистов и программистов, работающих в сфере компьютерной лингвистики, на эту конференцию приезжают и лингвисты, и инженеры корпусов.

Этот год – юбилейный для конференции. Первая состоялась в 1998 году в Гранаде, по инициативе Антонио Замполли, профессора из Института компьютерной лингвистики (Пиза, Италия). Прошло десять лет, и вот уже учреждена премия имени А. Замполли, одна из самых престижных в области компьютерной лингвистики. В этом году она была вручена Й. Уилксу из Университета Шеффилда (Великобритания). Лауреат сделал обзорный доклад о развитии корпусной лингвистики и теории автоматической обработки языковых данных за последние 45 лет. Все начиналось с новостных корпусов незначительного объема, затем была поставлена задача снятия неоднозначности, в настоящем развиваются диалоговые корпуса, а будущее, как считает Й. Уилкс, – за всемирной семантической сетью (Semantic Web) и всеобъемлющими моделями.

По словам председателя Оргкомитета конференции Н. Кальцолари (Институт компьютерной лингвистики, Пиза, Италия), корпусные методы уже не нужно защищать и продвигать: их эффективность говорит сама за себя. Н. Кальцолари подчеркивает, что конференция LREC – один из способов организовать интеграцию разных специалистов по языковым ресурсам и языковым технологиям.

С недавнего времени наличие конференции и ее ресурсов дополняется изданием одноименного журнала («Language resources and evaluation»), под редакцией Н. Иде и Н. Кальцолари.

Специалисты по письменной и устной речи, мультимодальности, исследователи терминологии, онтологии, собственно лингвисты, контент-провайдеры и другие профессиональные работники в этой области действуют независимо друг от друга. Все более разрастающееся поле исследований требует объединения и координации общих усилий.

Конференция состояла из основной части и семинаров; в основной части параллельно проходили доклады и стендовые сессии.

Были представлены несколько сотен лингвистических ресурсов – корпусов и систем автоматической обработки текста. Предлагались доклады, посвященные созданию и функционированию корпусов для исследования самых разных лингвистических задач: 1) корпуса звучащей речи, мультимодальные корпуса, 2) корпуса жестовых языков, 3) коллекции параллельных текстов на нескольких языках, 4) банки синтаксических деревьев, 5) корпуса с семантической разметкой, 6) корпуса кореферентности, 7) корпуса с размеченными именованными сущностями.

Многие ресурсы все еще находятся в стадии разработки и поэтому не могут пока считаться представительными. В то же время, были доклады и по крупным национальным проектам, например, по польскому, американскому и нидерландскому корпусам. Национальный корпус русского языка был представлен одним докладом про нестандартные формы в корпусе (Е. Гришина – Россия) и одним стендом про семантическую разметку в корпусе и снятие неоднозначности у имен существительных и прилагательных при помощи специальных правил – фильтров (О. Лашевская, О. Шеманаева – Россия).

Несколько заседаний были посвящены машинному переводу. Рассматривались задачи составления и предварительной обработки параллельных корпусов (М. Фишель, Х. Каалеп – Эстония), улучшения используемых статистических моделей языка (К. Лавеккиа и др. – Франция; М. Карпуа, Д. Ву – Гонконг; М. Карл – Германия), интеграции существующих языковых ресурсов (М. Итагаки, К. Аикава – Япония; Ю. Чен и др. – Германия; Б. Бабич и др. – Великобритания) и развития методов автоматической оценки качества перевода (Б. Бабич, Э. Хартли – Великобритания).

Было отмечено (С. Хасан, Х. Най – Германия), что современные технологии, находящиеся в открытом доступе, позволяют создать базовую систему машинного перевода

для произвольной пары языков всего за несколько недель. В то же время, дальнейшее улучшение и развитие подобных систем требует сложного лингвистического и математического моделирования.

В ходе дискуссий неоднократно высказывалось мнение, что дальнейший прогресс в области статистического машинного перевода невозможен без анализа сложных лингвистических структур. Это, в свою очередь, поднимает вопрос о применимости стандартных метрик автоматической оценки качества: BLEU, ROUGE и других. Б. Бабич и Э. Хартли представили исследование, демонстрирующее, что существующие метрики неприменимы для оценки машинного перевода нового поколения.

Тема использования Интернета в качестве источника языковых данных уже стала традиционной для LREC. В этом году, однако, она вызвала повышенный интерес в связи с пленарным докладом одного из основоположников теории информационного поиска Р. Баеса-Йейтса (США), который сопоставил различные виды знаний, представленные в Интернете. С одной стороны, пользователи ежедневно выкладывают в сеть мегабайты неструктурированной информации: тексты, картинки, видео и прочее. С другой стороны, появляется все больше успешных проектов по каталогизации и структуризации существующих знаний (например, веб-каталоги и Википедия). Р. Баеса-Йейтс оценил новейшие разработки в области семантической сети (Semantic Web) как уникальную возможность дать пользователям инструмент для самостоятельной структуризации своих данных. Такой подход выводит задачу интернет-поиска на принципиально новый уровень.

Структуризация и верификация представленной в Интернете информации становится все более актуальной и для компьютерной лингвистики. Если на предыдущих конференциях обсуждались, в основном, алгоритмы, основанные на частотах совместной встречаемости слов в Интернете, то сейчас все больше исследователей обращаются, например, к Википедии как к источнику языковых данных. Так, на конференции были представлены основанные на Википедии алгоритмы разрешения кореферентности (К. Мюллер и др. – Германия), составления лексико-семантических баз данных (Т. Цеш и др. – Германия) и онтологий (Г. Куи и др.– Гонконг).

Большинство исследований в области лексической семантики были посвящены онтологии WordNet и аналогичным проектам для других языков. Рассматривались способы автоматического пополнения подобных ресурсов с помощью статистического анализа семантической близости (Б. Брода и др. – Польша) и сопоставления уже существующих онтологий для других языков (Ф. Бонд и др. – Япония).

Ряд докладов был посвящен методам автоматического извлечения устойчивых сочетаний и терминов в различных языках: арабском (С. Булакнадель и др. – Франция), китайском (Ю. Янг и др. – Китай), латышском (Д. Дексне и др. – Латвия), словенском (С. Винтар, Д. Фишер – Словения) и других. Подобные алгоритмы позволяют расширять существующие онтологии за счет специальной лексики.

Для описания валентностных характеристик глаголов большинство исследователей используют теорию фреймов, предложенную Ч. Филлмором в конце 70–80-х годов. Был представлен целый ряд докладов о создании корпусов предложений с размеченной фреймовой структурой (аналоги американского проекта FrameNet) для различных европейских языков. FrameNet-ресурсы создаются на основе уже имеющихся корпусов с помощью дополнительного уровня разметки, что позволяет проследить взаимосвязь между различными слоями лингвистического описания. К сожалению, такая архитектура часто приводит к несоответствиям и противоречиям между разметками, что представляет большую проблему для создателей корпуса. Соответственно, несколько докладов были посвящены интеграции FrameNet-аннотаций в другие ресурсы: CCG корпус синтаксических деревьев (С. Боксвелл, М. Уайт – США) и онтологии (П. Воссен и др. – Нидерланды).

Два заседания и несколько стендовых докладов были посвящены синтаксису: созданию банков синтаксических деревьев и улучшению алгоритмов автоматического разбора предложения. Обсуждались проблемы синтаксической разметки в различных корпусах.

На нескольких заседаниях рассматривались проблемы дискурса, прежде всего, кореферентности. Были представлены корпусные исследования дискурсивных коннекторов (Э. Хинрикс, М. Лау – Германия) и несколько алгоритмов разрешения кореферентности для разных языков (И. Хендрикс и др. – Бельгия–Нидерланды; М. Мюллер и др. – Германия; К. Орасан и др. – Великобритания–Румыния–Португалия). О. Урюпина (Россия) проанализировала результаты работы современных систем разрешения кореферентности и отметила ряд подзадач, которые не могут быть успешно решены популярными в настоящее время статистическими методами. Во время последовавшей за до-

кладами дискуссии многие исследователи отмечали, что существующие корпуса кореферентности не могут считаться представительными. Соответственно, было предложено несколько новых проектов по созданию корпусов с дискурсивной разметкой.

Ряд докладов был посвящен созданию и поддержке корпусов устной речи. Рассматривалась взаимосвязь речи и возраста говорящего (С. Мёллер и др. – Германия–Великобритания), а также речевой ситуации (К. Марасек, Р. Губринович – Польша; Н. Моралес и др. – Испания). Кроме этого, были представлены корпуса, отражающие региональные особенности звучащей речи (К. Бринкманн и др. – Германия; Р. Мур – Австрия).

Наконец, одной из самых популярных тем конференции было взаимодействие человека и компьютера. Профессор Дж. Хиршберг (США) делала пленарный доклад о том, как распознавать автоматически, лжет человек или говорит правду. Исследователи этой проблематики обращали внимание на позы, поведение человека (например, трогает ли он волосы и лицо во время разговора), выражение лица, мимику, просодические характеристики, скорость сообщения, впечатление, которое производит рассказ и т.д. В исследовании Дж. Хиршберг параметры связываются с результатами тестирования испытуемых, одновременно со своим высказыванием тайно нажимающих педаль «правда» или «ложь». Их показания фиксируются и сопоставляются после опроса. При помощи этого психолингвистического теста был составлен небольшой корпус, который служит базой исследования речевых характеристик лжи. Проект, который впрочем пока находится в начальной стадии, преследует достаточно амбициозные цели – научить машину автоматически отличать ложь от правды.

На последовавших за докладом заседаниях были рассмотрены следующие практические задачи: 1) разработка диалоговой системы человек–машина, в которой по запросу человека выдается информация о наземном транспорте и его расписании; 2) разработка диалоговой системы человек – машина («умный дом») для пожилых людей; 3) разработка встроенной в автомобиль системы, распознающей голос пьяного человека (в Баварии собирают корпус звучащей речи для исследования особенностей речи людей с высокой дозой алкоголя в крови); 4) проекты по автоматическому извлечению информации и диалоговому общению человека с компьютером в музейной сфере: роботы, сконструированные таким образом, уже довольно успешно применяются в работе музеев.

Как мы уже говорили, на конференции было также проведено 25 семинаров и круглых столов по различным направлениям корпусных исследований. Особо отметим следующие темы:

- сбор материала и построение корпуса;
- лингвистическая аннотация корпусов;
- исследование эмоций и метафор в корпусе;
- мультимодальные корпуса;
- использование веб-пространства как корпуса;
- коллокации и устойчивые сочетания в корпусе;
- корпусное исследование диахронических изменений;
- корпуса звучащей речи;
- автоматическая обработка терминологии (особенно медико-биологической);
- корпуса языка глухонемых;
- автоматическая обработка малоисследованных языков.

На круглом столе, посвященном аннотации корпусов, были поставлены следующие задачи: 1) расширять круг корпусов и явлений, которые в них рассматриваются; 2) создавать корпуса разной модальности; 3) корпусные исследования должны предсказывать, какие области в языке наиболее подвержены изменению; 4) как и люди, машинные программы должны научиться автоматически извлекать только релевантную информацию.

В заключение обзора обратим внимание на принципиальные отличия LREC от уже упомянутой российской конференции «Диалог». На LREC почти полностью отсутствуют собственно теоретические доклады, не связанные с устройством корпуса, его статистической оценкой и его практическим применением. Напротив, область, которой было посвящено значительное количество докладов и под которую была даже выделена особая секция, – это проблема финансирования, поиск грантов, планирование проектов, спонсорство, отчетность и защита прав авторов материалов, из которых составляются корпуса. Главной же особенностью конференции LREC является разнообразие представленных проектов. По словам президента Европейской ассоциации языковых ресурсов (ELRA), если вы хотите пообщаться с каким-то конкретным специалистом в области языковых ресурсов и их оценки – приезжайте на эту конференцию, и он обязательно будет на ней выступать.

О.М. Урюпина, О.Ю. Шеманаева (Москва)