№ 1 2014

© 2014 г. С.Ю. ТОЛДОВА, О.Н. ЛЯШЕВСКАЯ

СОВРЕМЕННЫЕ ПРОБЛЕМЫ И ТЕНДЕНЦИИ КОМПЬЮТЕРНОЙ ЛИНГВИСТИКИ (в зеркале 24-й Международной конференции по компьютерной лингвистике COLING 2012, Мумбаи)

Статья посвящена обзору современных тематик и актуальных направлений компьютерной лингвистики на основе анализа материалов одной из конференций в этой области, а именно 24-й Международной конференции по компьютерной лингвистике COLING 2012. В обзоре приводится анализ основных подходов и проблемных точек в таких традиционных областях автоматической обработки текста, как автоматический морфологический и синтаксический анализ, машинный перевод и др. Также подробно рассматриваются современные задачи автоматического извлечения информации из текста, такие как извлечение фактов, извлечение мнений, анализ контента на основе привлечения онтологических ресурсов веба. Делается вывод о том, что для современного уровня развития компьютерной лингвистики характерно вовлечение все более сложных уровней лингвистического анализа в сферу автоматического анализа, применение гибридных подходов в решении задач компьютерной обработки текстов, совмещающих машинное обучение и алгоритмические методы. При этом уровни сложности современных задач обработки текстов, таких как извлечение временнóй референции в тексте, анализ структуры дискурса и многие другие, требуют активного привлечения экспертных лингвистических знаний.

Ключевые слова: компьютерная лингвистика, автоматический анализ текста, извлечение информации из текста, машинное обучение, гибридные методы, экспертная лингвистическая аннотация

This paper is an overview of the current issues and tendencies in computational linguistics. The overview is based on the materials of the conference on computational linguistics COLING 2012. The modern approaches to the traditional NLP domains such as postagging, syntactic parsing, machine translation are discussed. The highlight of automated information extraction, such as fact extraction, opinion mining are also in focus. The main tendency of modern technologies in computational linguistics is to accumulate the higher level of linguistic analysis (discourse analysis, cognitive modeling) in the models and to combine machine learning technologies with the algorithmic methods based on deep expert linguistic knowledge.

Keywords: computational linguistics, natural language processing, machine learning, expert linguistic annotation

ВВЕДЕНИЕ

Настоящая статья посвящена рассмотрению основных тенденций и направлений современной компьютерной лингвистики. Несмотря на то, что основная проблематика этой области прикладной лингвистики – алгоритмы обработки языковых данных и их использование в компьютерных приложениях – далека от традиционных тем журнала «Вопросы языкознания», мы надеемся, что наш обзор позволит читателям познакомиться с современными подходами к ряду «горячих тем» компьютерной лингвистики, а также с некоторыми особенностями методики и формата компьютерно-лингвистических исследований.

Прежде чем мы перейдем к рассмотрению этой тематики, хотелось бы остановиться на проблеме взаимодействия теоретической лингвистики и таких инженерных приложений, как обработка текста или извлечение информации из текста. Если начальный этап становления компьютерной лингвистики как самостоятельной научной отрасли (60-е годы прошлого столетия) характеризовался главенствующей ролью теоретических лингвистических моделей при решении прикладных задач, то в дальнейшем в зарубежных прикладных разработках фокус почти полностью сместился на применение математических статистических методов, разработку новых методов машинного обучения и их применение к языковой реальности на практике. В отечественной традиции теоретический подход существенно дольше сохранял свои позиции. Однако следует отметить, что это утверждение справедливо не для всех направлений обработки текстов. В задачах информационного поиска по мере развития технологий поиска по неструктурированным данным разработчики неизбежно были вынуждены обращаться к различным статистическим методам. Если же говорить о стандартных задачах собственно лингвистической обработки текста, таких как морфологический, синтаксический, семантический анализ, до сих пор во многих отечественных системах эти задачи решаются с использованием эвристических правил и базируются на теоретических лингвистических моделях.

Что касается зарубежных разработок, следует констатировать, что активно развивающиеся технологии компьютерной обработки большого объема данных (например, текстов разных стилей, жанров, тематики), доступность этих данных в компьютерном виде создали условия для проведения различных статистических экспериментов, для применения к лингвистическим задачам методов машинного обучения. Такое «статистическое» направление, получив быстрые и достаточно высокие результаты относительно простыми математическими методами, достигло некоторого своего порога. Существующие методы и технологии позволяют использовать универсальные математические модели для быстрого решения различных задач обработки и анализа текста. В результате на современном этапе развития компьютерной лингвистики помимо поиска новых еще более изысканных и интеллектуальных методов машинного обучения исследователи обратились к вовлечению в модели более сложных языковых фактов, к решению высокоуровневых лингвистических задач. О таком изменении тенденций можно судить по проблемам, обсуждаемым на различных крупных конференциях по компьютерной лингвистике, по тематике таких конференций.

В настоящей статье мы представляем обзор современных тематик с привлечением в качестве иллюстраций материалов только одной конференции, а именно 24-й Международной конференции по компьютерной лингвистике, состоявшейся 8–15 декабря 2012 г. в Мумбаи, Индия (http://coling2012-iitb.org/). Конференция проходит раз в два года под эгидой Международного комитета по компьютерной лингвистике (ICCL). В этой конференции приняло участие около 800 человек со всех континентов: сотрудников университетов, исследовательских центров, коммерческих корпораций – математиков, IT-специалистов, лингвистов и т.п.

Как уже отмечалось, одной из современных тенденций является повышенный интерес к теоретическим дисциплинам, так или иначе связанным с задачами компьютерной лингвистики. Так, на конференции немногочисленные пленарные доклады преследовали цель ознакомить компьютерных лингвистов с положением дел в смежных областях. С приглашенными докладами выступили директор Школы устного перевода Женевского университета Барбара Мозер-Мерсер (нейролингвистические исследования о приспособлении работы мозга к задачам профессиональных синхронных переводчиков) и бывший президент Национальной парламентской библиотеки Японии Макото Нагао (цифровые библиотеки и роль обработки естественного языка в их развитии).

Тематика двух других приглашенных докладов была связана со страной проведения конференции, Индией, и с индийской лингвистической традицией. Профессор Института информационных технологий в Хайдарабаде Дипти Мисра Шарма провела параллель между идеями школы Панини и практикой создания современных компьютерных

ресурсов, прежде всего для индийских языков и других языков с развитой морфологией и свободным порядком слов. Профессор Пол Кипарски также предлагает взглянуть на грамматику Панини с современных позиций, как на набор классификаций, правил и алгоритмов, на котором основывается компьютерная система. С этой точки зрения оказывается, что Панини предложил описание санкрита минимальной длины. В этом случае все особенности грамматики (грамматические категории и лексические классы, правила, их конкуренция и порядок применения, цикличность, блокировка, аналог тета-ролей, иерархии наследования и др.) служат для компрессии описания без потери точности и полноты – что, как нетрудно догадаться, представляет собой классическую задачу оптимизации работы компьютерно-лингвистического модуля. Кипарски задается также вопросом: может ли аналогичная идея минимизации длины метаданных лежать в основе деятельности человека при освоении языка? Его ответ – нет: на оптимизацию грамматики ушел труд нескольких поколений индийских грамматистов, в то время как человек усваивает язык слишком быстро.

Тематика конференции была весьма разнообразна и затрагивала практически все актуальные и активно развиваемые направления современной компьютерной лингвистики, начиная от исследования психолингвистических мотивов языкового поведения людей и заканчивая сложными математическими моделями машинного обучения без использования какого-либо предварительного лингвистического знания. Безусловно, невозможно в пределах одного небольшого обзора охватить все темы, обсуждавшиеся в рамках большой компьютерной конференции. Ниже мы остановимся на отдельных актуальных темах современной компьютерной лингвистики и кратко охарактеризуем решения, предлагавшиеся в докладах участников.

1. СОВРЕМЕННЫЕ ТЕНДЕНЦИИ В ОБЛАСТИ МАШИННОГО ПЕРЕВОДА

1.1. Общие направления исследований

Одной из доминант конференции стал машинный перевод. Этому направлению было посвящено несколько секций, а также многочисленные стендовые доклады. В последнее время потребность в автоматическом переводе чрезвычайно возросла, возрос и интерес исследователей к данной области компьютерной лингвистики. Это связано с активным развитием мультиязычной интернет-среды. С одной стороны, все больше документов в Сети представлено не на английском, а на других языках, все в большей степени растет потребность в интеграции знаний в различных областях деятельности на разных языках. С другой стороны, растет количество доступных в Сети текстов, представленных сразу на двух и более языках. Информация, извлекаемая из большого количества параллельных текстов, позволяет опираться в переводе не на экспертные оценки и правила, а на статистическую информацию. Эти обстоятельства обусловили развитие языковонезависимых подходов в области машинного перевода, а также дали толчок к активному использованию методов машинного обучения. Кроме того, наличие большого количества параллельных текстов в Сети, а также высокая «мобильность» лексических единиц – активное введение в оборот новой терминологии, изменение значений слов в контексте современных реалий – сделали возможным и необходимым автоматическое создание мультиязычных словарей. Таким образом, в рамках направления «машинный перевод» обсуждались следующие вопросы:

- автоматическое извлечение переводных эквивалентов из больших корпусов текстов;
- использование одноязычных корпусов при машинном обучении в системах машинного перевода;
- развитие семантических подходов в системах, основанных на статистических методах, включая интеграцию информации о предикатно-аргументной структуре предложения;

- разработка методов улучшения результатов работы систем, основанных на машинном обучении;
- проблемы анализа и перевода сложных слов в системах статистического машинного перевода;
- развитие методов измерения качества машинного перевода;
- анализ сложностей и методы их преодоления при машинном переводе для разноструктурных языков.

1.2. Автоматическое создание мультиязычных словарей

Технология создания мультиязычных словарей терминов и многословных терминов разрабатывается достаточно давно. Однако задача нахождения переводных эквивалентов в ситуации, когда одной лексеме соответствует целое словосочетание, представляет значительную сложность для статистических методов. Еще более сложной задачей является нахождение переводных эквивалентов морфологически сложных слов. Методам нахождения соответствий между сложными словами в языке-источнике и словосочетаний в целевом языке (ср, например, англ. *post-menopausal* vs. франц. *après la ménopause*) был посвящен доклад группы исследователей из Франции [Delpech et al. 2012]. Авторы предлагают композициональный подход к данной задаче. Слово на языке-источнике подвергается морфологическому анализу, далее находятся переводные эквиваленты его частей. Доклад интересен еще и тем, что для обучения системы авторы предлагают использовать близкие по тематике тексты на целевом языке, не являющиеся переводами. Обычно основные разработки ведутся на основе параллельных корпусов (в которых предложению на языке оригинала поставлено в соответствие предложение перевода). Однако в переводных текстах на выбор тех или иных языковых выражений оказывает влияние текст оригинала. Нередко в качестве перевода выбираются не самые идиоматичные выражения. Использование оригинальных текстов на целевом языке позволит найти более адекватные языковые выражения в качестве перевода. Проблема перевода сложных слов также освещалась в докладе совместного исследовательского коллектива из Оксфордского университета и Университета Карнеги-Меллон [Botha et al. 2012], который был посвящен переводу сложных слов с немецкого языка в статистических системах машинного перевода. В связи с тем, что компоненты сложных слов имеют относительно свободную сочетаемость при образовании этих слов, сложные слова часто не представлены в словарях и имеют низкую частоту в текстах, такие слова оказываются «не видны» в статистической модели, которая строится на основе этого корпуса. Для решения проблемы авторы доклада предлагают интегрировать различные вероятностные оценки отдельных компонентов сложного слова в статистическую модель. Обычно в статистической модели учитывается только частота сочетаемости синтаксически главного компонента сложного слова с другими словами из предложения. Авторы предлагают также учитывать различные вероятностные оценки для модифицирующего компонента.

1.3. Развитие гибридных подходов к машинному переводу

Отдельная секция была посвящена гибридным подходам к машинному переводу, совмещающим машинное обучение и правила, основанные на экспертных знаниях. Одной из сфер применения гибридного подхода является перевод в некоторой узкоспециальной области. В связи с тем, что статистические модели в системах машинного перевода базируются на данных параллельных текстов, качество таких систем зависит от объема и специфики доступных корпусов текстов. В узких областях такие параллельные корпуса не всегда доступны, в результате чего возникает так называемая проблема разреженности данных: некоторые цепочки лексем, особенно специфические терминологические словосочетания, встречаются в корпусе слишком редко, чтобы можно было надежно поставить им в соответствие переводные эквиваленты. Один из способов борь-

бы с такими ситуациями был представлен в докладе тайваньских ученых [Chen H.-B. et al. 2012]. Авторы представили новый метод получения данных для обучения. Обычно для перевода редких терминов используют специальные терминологические словари. Однако это не всегда помогает избежать ошибок, частота терминологического словосочетания в цепочке с некоторым общеупотребительным словом все равно оказывается низкой. Возникают ошибки с переводом общеупотребительной части такой цепочки. Например, при переводе цепочки «suffer from crystal induced arthritis» «редкость» самого терминологического словосочетания может помешать распознать словосочетания *suffer from* как фразового глагола. Для улучшения ситуации авторы предлагают подход, при котором вначале специфическая для некоторой предметной области часть текста подвергается упрощению – заменяется на более общеупотребительные аналогичные термины: например, *crystal induced arthritis* заменяется на *cancer*, *pneumonia* или *hypertension*. При такой замене система лучше справляется с переводом общеупотребительной части цепочки. Далее происходит обратная замена терминологической части на более специфическую с использованием специального терминологического словаря соответствующей предметной области. Проведенные авторами доклада эксперименты в области медицинских текстов показывают, что такой подход позволяет улучшить качество статистического машинного перевода.

1.4. Развитие семантических компонентов в статистических системах машинного перевода

Ориентация на семантический уровень анализа при машинном обучении обсуждалась в связи с несколькими задачами. Отдельной проблемой для статистического машинного перевода является определение переводных эквивалентов в случае служебных слов (функциональных слов), таких, например, как артикли, связки. Особенно это касается разноструктурных языков. Во-первых, позиции служебных слов (например, связок) в предложении могут не совпадать. Во-вторых, такие слова могут не иметь переводного соответствия на другом языке (например, артикли). При статистических подходах достаточно часто система ошибочно приписывает таким словам какие-либо переводные соответствия. В докладе исследователей из Киотского университета [Nakazawa, Kurohashi 2012] для адекватного перевода служебных слов было предложено учитывать в статистической модели синтаксические связи между знаменательным словом и служебным (в обсуждаемой системе используется формализм деревьев зависимости). Авторы предложили использовать связи от значимого слова к служебному. Проведенные авторами доклада эксперименты показали, что такой семантический, а не синтаксический принцип построения дерева дает лучшие результаты.

Значимость семантической информации и способ ее интеграции в статистических системах машинного перевода также обсуждалась в докладах международных исследовательских коллективов [Jones et al. 2012] (Великобритания, США), [Feng et al. 2012] (Германия, Китай) и др. В [Jones et al. 2012] предлагается метод статистического машинного перевода с использованием графического представления семантических отношений в тексте. В данном случае задачей машинного обучения является нахождение соответствий не между линейными фрагментами предложений на языке-источнике и на языке перевода, а соответствия между фрагментом и семантическим графом. Этот метод авторы доклада применили для перевода запросов к базам данных на естественном языке. Другое направление «семантизации» систем статистического машинного перевода – это использование информации о предикатно-аргументной структуре предложений, о семантических ролях именных групп в предикации. Так, например, в [Feng et al. 2012] представлены результаты эксперимента по использованию модуля приписывания семантических ролей в предложении (semantic role labeling). Как отметили в своем докладе авторы, с одной стороны, данная информация позволяет отслеживать так называемые «дальние» связи в предложении, с другой, позволяет более адекватно делить предложение на отдельные составляющие, учитываемые при поиске переводных

соответствий. В докладе группы ученых из Китайской академии наук [Zhai et al. 2012] обсуждался метод трансформаций предикатно-аргументной структуры, извлекаемой из предложения на языке оригинала (на основе приписывания семантических ролей в исходных предложениях), в соответствующую структуру целевого языка. Такой подход позволяет улучшить качество перевода за счет изменения порядка слов в переводе на основании информации о том, как устроена модель управления соответствующего предиката на языке перевода. Система включает три этапа. Первый этап – выделение моделей управления. На этом этапе приписываются семантические роли в тексте на языке оригинала. Второй этап – трансформационный: исходная предикатно-аргументная структура преобразуется в предикатно-аргументную структуру на языке перевода, для этого используются специальные трансформационные правила. На третьем этапе, этапе перевода, отдельно переводятся предикат и его аргументы, а затем с использованием специального алгоритма типа SKY осуществляется окончательный перевод. Эксперименты, проведенные авторами, показали существенное улучшение качества перевода при таком подходе.

1.5. Статистические системы постредактирования

Достаточно часто результат машинного перевода нуждается в дальнейшем постредактировании, независимо от того, основан он на правилах (RBMT – Rule-based machine translation) или на статистических моделях (SMT – Statistic machine translation). Постредактирование может также осуществляться с использованием машинного обучения (SPE – Statistical post-editing). Успешность машинного обучения в значительной степени зависит от того, какие данные берут разработчики системы в качестве обучающих. С одной стороны, обучающий корпус может представлять собой результат ручного редактирования ошибок машинного перевода. В таком случае система тренируется только исправлять ошибки, сделанные на первом этапе перевода. С другой стороны, в качестве эталона можно взять уже готовый перевод и обучать систему исправлять перевод, полученный на первом этапе, до состояния эталона. В этом случае система будет ориентироваться не собственно на ошибки, а на расхождения между первичным переводом и эталоном. При этом источником расхождения между первичным переводом и эталоном может быть вовсе не ошибка системы, а возможность синонимичных перифраз. Сравнение разных подходов к обучению в статистических системах постредактирования обсуждалось в докладе группы исследователей из Ирландии, США и других стран [Béchara et al. 2012]. Исследователей интересовал подход к обучению, при котором эталонным корпусом является готовый текст на целевом языке. Подготовка таких обучающих корпусов менее трудоемка и не требует дополнительных человеческих ресурсов. Качество машинного перевода также может быть оценено разными способами: автоматически (с использованием разных метрик близости результата к эталонному переводу) и экспертно (качество перевода оценивается людьми). Результаты исследования показали, что если при применении автоматических метрик выше оцениваются результаты работы статистических систем МП без постредактирования, то экспертные оценки выше для систем, основанных на правилах, со статистическим постредактированием.

1.6. Проблемы перевода разноструктурных языков

Многие доклады, представленные на секциях по машинному переводу, были посвящены переводной паре языков китайский-английский (см., например, работу [Zhai et al. 2012], обсуждаемую в п. 1.4, [Xiao X. et al. 2012; Gong et al. 2012] и др.). Сложности, возникающие при автоматическом переводе для данной пары языков, связаны с тем, что они очень сильно различаются по структуре, порядку слов, способам выражения грамматических категорий. Так, доклад исследователей из Сингапура и Китая [Gong et al. 2012] был посвящен проблеме правильного выбора временно́го оформления глагола при условии, что в системе языка оригинала, а именно в китайском языке, отсутствует

грамматический способ выражения времени. Проблемы, связанные с разными способами выражения грамматических категорий, возникают и при английско-японском переводе. В докладе ученых из Киотского университета [Nakazawa, Kurohashi 2012] (обсуждение см. в п. 1.4) было показано, каким образом можно решать вопрос о расхождениях в составе и синтаксическом поведении служебных слов. В рамках данного направления обсуждались также проблемы машинного перевода для языков, для которых недостаточно лингвистических ресурсов или ресурсы совсем отсутствуют, при разработке систем автоматического анализа текста. Это, например, диалекты арабского языка (см. работу исследователей из Колумбийского университета [Salloum, Elissa 2012]).

1.7. Методы оценки качества перевода

В последнее время в связи с более широким использованием систем машинного перевода достаточно остро стоит вопрос об оценке качества перевода. Даже при экспертной оценке того, какой из переводов предложения-источника более адекватен, мнения экспертов часто расходятся. Процедура ручной оценки слишком трудоемкая, и не всегда возможно ее осуществить. Возникает задача определения некоторой формальной метрики оценки качества перевода, а также задача автоматического ранжирования нескольких вариантов предложенных машиной переводов одного и того же исходного предложения по качеству. Такому автоматическому «оценщику» был посвящен доклад ученого Немецкого исследовательского центра искусственного интеллекта Э. Аврамидиса [Avramidis 2012]. В работе представлен автоматический классификатор, который обучается на переводах, ранжированных вручную, а также с учетом ряда лингвистических критериев.

2. СОЗДАНИЕ ГЛУБОКО АННОТИРОВАННЫХ ЛИНГВИСТИЧЕСКИХ РЕСУРСОВ

Активное использование методов машинного обучения в системах автоматической обработки текста, постоянное усовершенствование этих методов служит стимулом для разработки и создания новых лингвистических ресурсов. Развитию таких ресурсов, проблемам, связанным с их аннотированием, было посвящено немало докладов. Следует отметить, что одной из основных тенденций является разработка аннотаций более высоких языковых уровней, таких, как анафорическая разметка, дискурсивная разметка и др. В то же время пополняются и обновляются уже существующие ресурсы. По материалам конференции можно обозначить следующие активно развивающиеся направления в данной области:

- развитие и усовершенствование существующих корпусов, особенно в области синтаксического аннотирования;
- создание корпусов со специфическими видами разметки, такими как, например, разметка локативных отношений, временно́й референции (данные виды разметки актуальны при разработке систем извлечения информации из текста);
- разработка специальных интерфейсов для разметки корпусов и методов тестирования ошибок разметки.

2.1. Усовершенствование существующей разметки корпусов и развитие новых типов аннотаций

Проблемы усовершенствования разметки в корпусах были отражены в целом ряде докладов. Отдельно следует остановиться на докладе чешских ученых из Карлова университета в Праге. Доклад был посвящен Пражскому корпусу (Prague Dependency Treebank, PDT) с синтаксической разметкой в терминах деревьев зависимостей [Bejček et al. 2012]. Этот корпус представляет интерес, поскольку он может служить эталоном для синтаксически размеченных корпусов славянских языков. На конференции был представлен новый релиз корпуса PDT 2.5, в котором появились новые типы разметки.

Предыдущая версия представляла собой коллекцию газетных текстов начиная с 1990-х гг. издания, снабженных морфологической разметкой, поверхностно-синтаксической разметкой (отражающей базовые синтаксические отношения, такие как подлежащее, дополнение и т.п.), тектограмматической разметкой. На уровне тектограмматической разметки элементами разметки являются уже не просто отдельные словоформы. На этом уровне признаки приписываются либо значимым словам, либо сочетанию служебных и значимых слов (например, паре предлог – существительное). Аннотация данного уровня включает разметку семантических ролей, таких как актор, адресат и т. п., разметку анафорических связей, разметку, отражающую тема-рематическое членение предложения. PDT включает 7110 документов, размеченных вручную, что составляет 115844 предложения или 1957247 словоупотреблений. На текстограмматическом уровне размечено 50 % текстов. Новая версия корпуса включает также три новых типа разметки: во-первых, разметку устойчивых словосочетаний, включая именованные сущности, такие как персоны, географические названия и др., во-вторых, разметку специфических значений множественного числа конкретно-референтных существительных: множественное число с семантикой парности (например, словоформа *руки* по умолчанию имеет не ‘количество рук’, больше одной, а ‘две руки’), групповое множественное (ср. значение словоформы *ключи* – ‘связка ключей’), в-третьих, разбиение сложных предложений на предикации.

Как было отмечено выше, в настоящее время создается достаточно много ресурсов со специальными типами разметки. Так, например, доклад [Coyne et al. 2012] был посвящен аннотации корпуса специальными 3D-иллюстрациями. Глобальной задачей авторов доклада является моделирование преобразования текста в изображения. В докладе авторы представляют систему аннотирования фреймов FrameNet, описывающих некоторую ситуацию реального мира, условными 3D-изображениями, иллюстрирующими данные фреймы. Авторов интересуют фреймы, в которых связываются предикаты и обозначения местоположений. Они отмечают, что для реализации такой задачи необходимо понять, как соотносятся реальные местоположения и их языковое выражение.

Другой тип прагматической информации, в терминах которой производится аннотирование, – это временна́я референция. В докладе [Bejček et al. 2012] анализируется опыт коллективной разметки корпуса (так называемой краудсорсинговой разметки, англ. crowdsourcing, подробнее см. п. 8.2). Обсуждаемая в докладе разметка отражает временное отношение события, выраженного глаголом в некоторой видо-временной форме, и языкового выражения, имеющего временную референцию, например, *в понедельник* или *вчера*. Исследование показало, что возможно обеспечить достаточно точную разметку силами неспециалистов, используя только минимальный набор простых, надежно определяемых признаков. Это позволит быстро увеличить объем данных, необходимых для машинного обучения. Предлагается также учитывать результаты синтаксического анализа, а именно степень подобия синтаксических деревьев (деревьев зависимости): подобия путей по дереву от языковых выражений, имеющих временную референцию, до предикатов, обозначающих соответствующее событие. Как показывают эксперименты, проведенные авторами доклада, это позволяет повысить точность разметки.

2.2. Развитие инструментов и методов разметки и контроля качества

В связи с тем, что объем аннотируемых корпусов все время увеличивается, для разметки корпусов активно применяется автоматическое аннотирование, усложняются сами параметры, по которым происходит аннотирование, актуальной становится проблема ошибок как при автоматической, так и при ручной разметке. Данному направлению был посвящен ряд докладов в рамках секции по аннотированным ресурсам, а также целый ряд докладов постерной секции. Дело в том, что даже при ручной разметке корпуса возникают ошибки и несогласованность в аннотациях. Достаточно часто это не просто случайные произвольные ошибки, а ошибки, связанные с определенными

типами сложностей и переходных случаев. Такие ошибки влекут в дальнейшем ошибки в машинном обучении. Использование корпуса с большим количеством ошибок как «золотого стандарта» влечет неправильную оценку качества работы систем. Таким образом, важно научиться выявлять регулярные типы ошибок и оценивать их последствия. В совместном докладе исследователей из разных научных коллективов Франции [Mathet et al. 2012] обсуждается метод моделирования различных классов ошибок, совершаемых аннотаторами, а также специальный инструмент для применения данного метода. С помощью предлагаемого инструмента можно оценивать степень влияния разных типов ошибок на общее качество разметки. Еще один доклад французских исследователей [Fort et al. 2012] был посвящен проблемам ручной разметки корпусов. Ручная разметка является достаточно трудоемкой задачей. Она требует тщательной подготовки, включающей формулировку задачи, подготовку исходных данных, подбор аннотаторов, разработку схемы аннотации, а также разработку системы контроля качества разметки. При организации процесса разметки необходимо заранее оценить степень сложности разметки корпуса по соответствующим лингвистическим признакам, а также по возможности предложить некоторую «декомпозицию» сложной задачи на более простые. Авторы, анализируя различный опыт аннотирования корпусов, описанный в литературе, предлагают некоторую шкалу оценки сложности задачи с точки зрения разных аспектов разметки, которая позволила бы заранее оценить трудоемкость организуемой процедуры аннотирования.

3. КОМПЬЮТЕРНАЯ ЛЕКСИКОГРАФИЯ

Достаточно широко на конференции была представлена такая область компьютерной лингвистики, как компьютерная лексикография. Эта область имеет давнюю традицию. Одно из основных направлений исследований в данной области – статистические методы автоматического извлечения семантических отношений между лексемами. Статистический подход в лексикографии базируется на представлении о том, что семантическая связанность двух или более слов отражается в частоте их совместной встречаемости. В зависимости от задачи могут варьироваться ограничения на то, какие слова в тексте подпадают под рассмотрение, какие статистические методы используются для определения семантических отношений между лексемами или же для определения конкретного значения лексемы в контексте, а также какие дополнительные источники (лексикографические, энциклопедические, текстовые) привлекаются для семантического анализа.

3.1. Методы разрешения семантической неоднозначности

Одной из наиболее обсуждаемых и активно разрабатываемых тем является разрешение семантической неоднозначности (word sense disambiguation). Данной тематике было посвящено несколько заседаний специальной секции. Разрешение семантической неоднозначности является актуальной задачей для различных приложений по извлечению информации из текста, в частности, в приложениях по извлечению именованных сущностей. Разработка автоматических методов разрешения семантической неоднозначности ведется уже более двадцати лет. Однако до сих пор данная задача остается актуальной.

Одним из первых методов автоматического разрешения семантической неоднозначности является алгоритм Леска [Lesk 1986]. Этот метод предполагает вычисление значения некоторой лексемы в конкретном контексте на основе пересечения множества слов из словарного толкования соответствующей лексемы и контекстных лексем. Например, для словосочетания *pine cone* значение лексемы *pine* может быть вычислено на основе того, что в одном из толкований присутствует лексема *tree*, эта же лексема присутствует в соответствующем толковании лексемы *cone*. В результате можно сделать вывод, что в данном контексте имеется в виду *pine* в значении ‘дерево’. Данный метод обладает целым рядом недостатков. В частности, далеко не всегда толкования двух контекстно

близких лексем имеют пересечения. Многие современные исследования представляют собой развитие и усовершенствование данного алгоритма. В более общем виде метод пересечения контекстов при разрешении неоднозначности можно переформулировать следующим образом: каждому значению лексемы можно поставить в соответствие некоторый список диагностических лексем; соответствующее значение некоторой лексемы в контексте вычисляется при помощи пересечения диагностического множества данной лексемы и диагностического множества лексемы, встретившейся в одном контексте с данной. Одно из направлений усовершенствования – это привлечение дополнительных интернет-ресурсов для создания диагностических списков, таких, как, например, WordNet[1], Wikipedia[2], Wiktionary[3]. Другой путь разрешения неоднозначности, обсуждавшийся в рамках конференции, – это использование различных методов группировки лексем на основе статистического анализа корпусов текстов большого объема. Доклады, представленные на конференции, были посвящены различным методам оптимизации использования вышеперечисленных ресурсов. Так, доклад группы ученых из США [Shen et al. 2012] был связан с использованием данных Wikipedia. Различение омонимичных терминов в Википедии обеспечивается системой ссылок на соответствующие страницы. Ссылки снабжены соответствующими «ярлыками» (ср., например, «ключ, родник», «музыкальный ключ» и т.п.). Такие ярлыки активно используются в автоматических системах разрешения неоднозначности. Обсуждаемый доклад посвящен анализу различных способов организации таких ссылок. В результате авторам доклада удалось выявить случаи непоследовательного приписывания тегов в Википедии, которые могут приводить к ошибкам при автоматическом использовании этого ресурса для разрешения неоднозначности, а также предложить некоторые способы преодоления таких непоследовательностей.

В докладе немецких ученых [Meyer, Gurevych 2012] обсуждались преимущества использования такого ресурса, как Wiktionary. Во-первых, этот словарь охватывает больший объем лексических единиц по сравнению с обычными ресурсами, создаваемыми экспертами-лексикографами, типа WordNet. При этом извлекаемые на основе этого ресурса списки близких лексем содержат существенно меньше ошибок, чем списки, извлекаемые статистически на основе анализа больших корпусов. В-третьих, если на основе Wikipedia в основном можно извлекать необходимую информацию о существительных, то данный ресурс позволяет вычислять степень семантической близости между глаголами. В-четвертых, это мультиязычный словарь. Соответственно, возможно использовать информацию о переводных эквивалентах.

Авторы доклада [Miller et al. 2012] предлагают пополнять диагностические списки контекстов за счет автоматического анализа корпуса большого объема и создания так называемого дистрибуционного тезауруса. Диагностический список для каждой лексемы строится на основе статистической оценки ее сочетаемости с учетом типа синтаксической связи.

Альтернативным источником информации для разрешения неоднозначности могут служить параллельные тексты на разных языках. Использование при этом так называемого латентно-семантического анализа позволяет группировать лексику по семантической близости, см., например, доклад другой исследовательской группы из Германии [Kim et al. 2012].

На конференции обсуждались как уточнения к традиционным статистическим методам, позволяющие улучшить качество статистических алгоритмов, так и новые методы, позволяющие автоматически разбивать лексику на семантические классы без участия человека. Так в работе французских исследователей [Schwab et al. 2012] предлагается использование так называемого муравьиного алгоритма.

---

[1] Электронный тезаурус / семантическая сеть для английского языка, разработанный в Принстонском университете URL: http://ru.wikipedia.org/wiki/WordNet

[2] http://en.wikipedia.org/wiki/Wikipedia:About

[3] Викислова́рь (англ. *Wiktionary*) – свободно пополняемый многофункциональный многоязычный словарь и тезаурус, основанный на вики-движке. Один из проектов фонда «Викимедиа». Сначала появился на английском языке 12 декабря 2002 года. URL: http://www.wiktionary.org/

3.2. Создание лексикографических ресурсов

Второе важное направление компьютерной лексикографии – это создание специализированных лексикографических ресурсов, в частности, лексических тезаурусов типа WordNet для разных языков. Здесь одной из теоретических проблем является «переносимость» семантических отношений, устанавливаемых между лексемами в англоязычном Принстонском WordNet [http://wordnetweb.princeton.edu/perl/webwn], на материал другого языка, применимость тех же принципов, а также возможность установления связей между единицами Принстонского WordNet и единицами тезауруса на описываемом языке. Этому вопросу частично был посвящен постер разработчиков польского WordNet-a [Rudnicka et al. 2012]. Интерес представляет и разработанный авторами доклада графический интерфейс для работы с семантической сетью WordNet-a, удобный для редактирования данных.

4. АВТОМАТИЧЕСКИЙ МОРФОЛОГИЧЕСКИЙ И СИНТАКСИЧЕСКИЙ АНАЛИЗ

Пожалуй, не менее значимым направлением по сравнению с машинным переводом, широко представленным в докладах на конференции, был автоматический морфологический и синтаксический анализ (парсинг).

Необходимо отметить следующие особенности современного состояния этой области автоматической обработки текста: во-первых, основные активно развиваемые и применяемые технологии морфологического и синтаксического анализа – это технологии машинного обучения. Во-вторых, можно констатировать, что современная компьютерная лингвистика достигла очень высокого уровня в этой области. Существует некоторое множество стандартных технологий машинного обучения, относительно легко переносимых на материал новых языков. Эти технологии позволяют сравнительно быстро получать неплохие результаты для языков самой разной структуры. В открытом доступе представлены автоматические анализаторы для очень многих языков. В-третьих, уровень развития базовых статистических технологий анализа достиг некоторого порога. Основные усилия исследований направлены на усовершенствование и «доработку» существующих методов машинного обучения. Актуальными направлениями в этой области являются исследование возможностей методов машинного обучения без учителя (извлечение морфологической и синтаксической информации из корпусов текстов большого объема, не имеющих предварительной лингвистической разметки), а также адаптация существующих методов к анализу менее разработанных в области автоматического анализа языков, например, языков с богатой флективной морфологией.

4.1. Автоматический морфологический анализ

Что касается автоматического морфологического анализа (pos-tagging), то усилия исследований на данный момент направлены, с одной стороны, на некоторые улучшения, различные вариации существующих методов машинного обучения в применении к морфологическому анализу. Проблемам развития таких методов были посвящены доклады [Waszczuk 2012; Zhao, Marcus 2012; Billingsley, Curran 2012] и др. Для применения наиболее распространенных методов, так называемых методов машинного обучения с учителем, требуются эталонные корпуса большого объема, имеющие эталонную морфологическую разметку. Создание таких обучающих корпусов является трудоемкой задачей. При этом оказывается, что результат обучения очень плохо переносится с одной предметной области на другую. Если анализатор обучался на материале новостных текстов, то качество его применения на материале медицинских текстов резко понизится. Разработка методов адаптации анализатора для разных предметных областей группы из Темпльского университета штата Филадельфия рассматривалась в работе [Xiao M. et al. 2012]. Другой выход из положения – это усиление существующих методов за счет

развития методов обучения без учителя. Этот подход обсуждался в работе группы исследователей из Пенсильванского университета [Zhao, Marcus 2012].

Кроме того, до сих пор статистический морфологический анализ представляет некоторую проблему для языков с развитой флективной морфологией, например, для многих славянских языков. Вопросам статистического автоматического анализа польского языка был посвящен доклад польских ученых [Waszczuk 2012].

С другой стороны, интерес представляют доклады, посвященные различным нестандартным задачам морфологического анализа. Так в работе [Waszczuk 2012] обсуждалась проблема деления на слова для языков, где невозможна опора на пробелы в качестве маркера границы слов, а именно для китайского языка. Доклад международной группы исследователей [Attia et al. 2012] был посвящен проблемам «восстановления» исходных форм (лемматизации) отсутствующих в словаре лексем для такого морфологически сложного языка, как арабский. Особым предметом исследования при разработке систем автоматического анализа текста являются биомедицинские тексты, которые обладают своей спецификой не только в области лексического состава, но и на более низких языковых уровнях. Так, например, для текстов данной предметной области необходимым компонентом морфологического анализа является анализ внутренней структуры поликорневых медицинских терминов (см. доклад французского исследователя [Claveau 2012]).

4.2. Автоматический синтаксический анализ (parsing)

Большое количество докладов, постеров и демонстраций были посвящены автоматическому синтаксическому анализу (так называемому парсингу). Здесь необходимо отметить две особенности. Во-первых, несмотря на то, что в теоретических работах грамматика непосредственных составляющих безусловно занимает лидирующие позиции (хотя такие формализмы, как HPSG и LFG, также популярны), в системах автоматического синтаксического анализа использование деревьев зависимости получает все большее распространение. Анализу в терминах деревьев зависимости было посвящено немало докладов, в том числе [Goldberg, Nivre 2012; Xiao M. et al. 2012] (см. также п. 4.1), [Agic 2012; Inokuchi, Yamaoka 2012] и др. Одним из актуальных методов синтаксического анализа, позволяющих разрешать синтаксическую омонимию, является «обогащение» вероятностных синтаксических парсеров информацией о лексических вероятностях (о лексической сочетаемости, об устойчивых словосочетаниях и т.п.). Методы «усиления» парсера, основанного на формализме непосредственных составляющих, информацией о лексической сочетаемости обсуждались в докладе ученых из Северо-восточного университета Китая [Zhu et al. 2012]. Еще одним из перспективных методов является «скрещивание» парсеров, использующих разные формализмы. Например, в докладе исследователей из Штутгарта [Farkas, Bohnet 2012] предлагалось при статистическом обучении использовать как признаки, необходимые для анализа в терминах непосредственных составляющих, так и признаки, значимые для анализа в терминах деревьев зависимости. В этом докладе на материале английского языка было показано, что совмещение двух моделей дает значимое повышение точности синтаксического разбора.

Современные методы машинного обучения базируются на обучающих корпусах большого объема. Многие обсуждения, посвященные синтаксическому анализу, затрагивали проблемы работы с эталонным корпусом. В частности, обсуждались проблемы установления соответствий наблюдаемых в реальных текстах синтаксических явлений с данными, получаемыми по размеченным корпусам, а также проблемы выявления ошибок в корпусной разметке (например, в постерном докладе исследовательской группы из Барселоны [Mille et al. 2012]). Постерный доклад исследователей из России [Gareyshina et al. 2012] был посвящен проведению форума по оценке методов автоматического синтаксического анализа для русского языка, в нем также приводилось сравнение результатов работы разных парсеров.

Приз за лучший доклад COLING-2012 выиграла команда, состоящая из двух PhD аспирантов американских университетов и их молодого научного руководителя [Nguyen

et al. 2012]. Их доклад был посвящен проблеме пересчета существующих аннотаций в другие формализмы (reannotation), а именно был представлен алгоритм перевода разметки составляющих Penn Treebank в формализм категориальных грамматик типа HPSG, который позволяет обнаружить вынесенные предикатно-аргументные зависимости (подъем правого аргумента, вынос wh-вопроса, сочинение, субъектный и объектный контроль и т.п.). Как нам представляется, выбор данного доклада показателен в нескольких отношениях. Во-первых, он показывает повышенный интерес к грамматике зависимостей, о чем уже было сказано выше. Во-вторых, доклад отражает «модную» в компьютерной лингвистике тенденцию, когда старые проверенные ресурсы получают «вторую жизнь» и могут быть использованы в задачах, для решения которых они по своей природе не предназначались. В-третьих, обращает на себя внимание своеобразный «конформизм» компьютерных лингвистов – готовность абстрагироваться от одного традиционного формализма, разобраться в другом, более подходящем для решения конкретной задачи, и представить данные в «реинкарнированном» виде.

4.3. Анализ «малоресурсных» языков (underresourced languages)

Вполне ожидаемо, на конференции в Азии была представлена большая доля докладов ученых из Китая, Японии, Южной Кореи, самой Индии, Ближнего Востока и Австралии. (Заметим, кстати, любопытную деталь: во многих докладах китайских и индийских университетов «третьим соавтором» участвовали профессора из Европы и Северной Америки – это очевидным образом помогло вывести оригинальные исследования на международную орбиту и качественно улучшить структуру и язык презентации.) Вместе с исследованиями, выполненными в Восточном полушарии, пришла и тематика автоматической обработки местных языков – индийских, австралийских, амхарского и т.п., которые менее прочих описаны в рамках компьютерных приложений. В частности, было представлено значительное число докладов и постеров по созданию электронных ресурсов для этих языков: обсуждение самих корпусов, проблем автоматического анализа, возникающих при аннотировании этих корпусов, методик обнаружения ошибок и несоответствий в аннотировании (см., например, исследование группы из Университета Страны Басков [Atutxa et al. 2012]). В связи с современными технологиями работы с мультиязычными ресурсами решение данных задач ведет к развитию нового подхода: быстрые решения на основе обработки мультиязычных параллельных корпусов, извлечение знаний о малоописанном языке на основе анализа данных таких корпусов.

5. АВТОМАТИЧЕСКИЙ АНАЛИЗ НА УРОВНЕ ДИСКУРСА

Современные задачи извлечения информации из неструктурированных текстов требуют выхода за рамки одного предложения. Возникает необходимость прослеживать различные отношения между объектами и событиями на уровне всего текста. Помимо исследований в области традиционных задач автоматической обработки текста на всех собственно языковых уровнях: на уровне морфологии, синтаксиса и семантики, – в последнее время активно развивается направление автоматического анализа более высоких уровней, а именно таких прагматических (дискурсивных) уровней, как разрешение кореферентности, анализ временно́й референции, анализ дискурсивных отношений. Ниже более подробно остановимся на рассмотрении представленных на конференции работ по автоматическому установлению кореферентных связей и дискурсивных отношений.

5.1. Автоматическое разрешение кореферентности

Одним из важных механизмов обеспечения связности текста является механизм поддержания референции. При извлечении информации из текста необходимо проследить упоминания одного и того же объекта. Это важно как для определения фокуса внимания текста (например, для определения наиболее значимого медийного объекта

в новости), так и для извлечения информации об участниках событий. При этом задача распознавания референциальных цепочек в тексте (последовательности именных групп в тексте, относящихся к одному и тому же референту или событию) может сводиться к задачам самых разных уровней сложности в зависимости от того, какие именно явления моделируются данным компонентом. Большинство систем разрешения анафоры имеют дело со случаями, когда референтом именной группы (местоимения) является конкретный объект действительности. В этой области достигнуты неплохие результаты.

В современных исследованиях наблюдается интерес как к усложнению самой задачи, так и к усложнению типов используемой информации, привлечению информации более глубинного уровня анализа текста. Ниже остановимся на двух докладах, хорошо иллюстрирующих данную тенденцию.

Более сложной задачей по сравнению со случаем установления анафорических связей для конкретно-референтных объектов является разрешение событийной анафоры – случаев, когда местоимение относится к некоторому событию в тексте. Если в случае конкретно-референтной анафоры, как правило, можно предъявить фрагмент (именную группу) из предшествующего контекста, к которому относится некоторое анафорическое местоимение, то в случае события это сделать достаточно сложно, а иногда и невозможно. В первом случае задача может быть решена с использованием минимального поверхностного лингвистического анализа. Так, многие успешные системы «останавливаются» на синтаксическом уровне. Как показали исследователи из Китая и Сингапура Фан Кун и Годун Чжоу в своем докладе [Kong, Zhou 2012], в случае событийной анафоры необходимо привлекать результаты более глубинного уровня анализа. Авторы рассматривают систему, в которой используется как локальная семантическая информация – информация о предикатно-аргументной структуре предложения, извлекаемая при помощи поверхностного семантического анализатора (shallow semantic parser), так и глобальная семантическая информация. Последняя представляет собой информацию о кореферентных связях именных групп с учетом ролей этих именных групп в аргументной структуре предложений из контекста.

В докладе исследователей из Японии и США [Inoue et al. 2012] рассматривается система распознавания кореферентности, основанная на логической системе построения и выбора гипотез, так называемом методе абдукций. В качестве исходной информации для построения различных гипотез о совпадении / несовпадении референтов двух именных групп авторы используют информацию о предикатно-аргументной структуре предложений, в которых эти именные группы встречаются. Так, некоторые предикаты или отношения между предикатами имплицируют несовпадение референтов, например, разные актанты одного глагола скорее не совпадают (если один из них не выражен возвратным местоимением). Другие же, наоборот, с большой вероятностью имплицируют совпадение референтов: например, именные группы с одинаковой семантической ролью при синонимичных предикатах, обозначающих одно и то же событие. Другой тип используемой в системе информации – это информация о семантических отношениях между именными группами: являются ли они синонимами, гиперонимами или же, наоборот, антонимами или когипонимами. Для вычисления семантических отношений авторы используют такой тезаурусный ресурс, как WordNet. Также в системе используются так называемые нарративные цепочки: определенные типы последовательностей предикаций предсказывают определенный тип соотношения референтов (ср., например, последовательность событий ‘толкнуть’ и ‘упасть’).

5.2. Дискурсивный анализ

На конференции также были представлены работы по автоматическому дискурсивному анализу. Обсуждались такие вопросы, как автоматическая оценка дискурсивной связности, использование глубокого синтаксического и тема-рематического анализа при распознавании дискурсивных отношений между предикациями на материале Пражского трибанка (см. доклад представителей Токийского технологического института [Iida,

Tokunaga 2012]), разрешение неоднозначности дискурсивных коннекторов с использованием параллельных корпусов для языков, в которых отсутствуют корпуса с дискурсивной разметкой (см. доклад исследователей из Гонконга и Катара [Zhou et al. 2012]), а также ряд других вопросов, связанных с аннотированием корпусов на уровне дискурса и с использованием этой разметки (см., например, [Mírovský et al. 2012].

6. СОВРЕМЕННЫЕ ТЕНДЕНЦИИ В ОБЛАСТИ ИНФОРМАЦИОННОГО ПОИСКА И АВТОМАТИЧЕСКОГО ИЗВЛЕЧЕНИЯ ИНФОРМАЦИИ ИЗ ТЕКСТОВ

В рамках конференции обсуждались также традиционные задачи таких направлений анализа контента, как информационный поиск, вопросно-ответные системы, извлечение информации из текста. Следует особо отметить некоторые современные тенденции в этой области:

- «семантизация» методов извлечения именованных сущностей и фактов (событий) из текста;
- активное развитие различных задач и методов распознавания тональности и анализа мнений (sentiment analysis и opinion mining);
- анализ социальных сетей и в особенности микроблогов;
- использование различных ресурсов веба, включая такие веб-онтологии, как Википедия или DBPedia и др. в различных задачах автоматической обработки контента, а также развитие семантического веба (W3C или semantic web), ресурсов открытых связанных данных (linked open data).

6.1. Извлечение именованных сущностей и извлечение событий из текста

Стандартной задачей в области извлечения именованных сущностей (named entities recognition, NER) является распознавание объектов трех основных классов: персон, организаций и местоположений. Основные методы, которые применяются в данной области, – это различные классификационные методы машинного обучения: на основе корпуса, размеченного вручную, система обучается распознавать, является ли некоторая цепочка слов объектом определенного класса или нет. Каждый из кандидатов первоначально получает описание в терминах некоторого множества признаков. Задача системы – на основе соотношения значений признаков принять решение о принадлежности кандидата к соответствующему классу объектов. В простом случае признаками для некоторой цепочки-кандидата являются контекстные слова. Например, признаком может служить существительное *футболист* для лексемы *Петров* в контексте *футболист Петров.* Лексема *футболист* является хорошим диагностическим признаком для персон. Качество системы в значительной степени зависит от того, какая информация включается в исходное множество признаков. Задача усложняется, если необходимо извлекать сущности других классов, например, различные продукты, фильмы и т.п. Также сложной задачей, до сих пор не имеющей общепринятого адекватного решения, является однозначная идентификация объекта в тексте в случае так называемой онтологической омонимии, когда совпадают наименования двух объектов (например, *Иванов Александр Александрович* является фамилией, именем и отчеством сразу нескольких людей, в частности астронома и шахматиста).

Основные направления развития методов в данной области – это расширение и детализация классов именованных сущностей, которые может распознавать система, а также «обогащение» множества признаков дополнительной семантической или онтологической информацией для разрешения онтологической омонимии.

Для более полного извлечения информации из текста достаточно часто необходимо распознавать более точно онтологический подкласс извлекаемой сущности. Так, доклад ученых из Пекинского университета [Li et al. 2012] был посвящен распознаванию различных подклассов персон. Такая задача является более сложной для реализации. Во-первых, самих выделяемых классов оказывается больше. Во-вторых, объекты в вы-

деляемых подклассах достаточно близки по своим семантическим свойствам. В задачи авторов входило распознавать персоны по роду занятия. Нередко контекстные слова не могут быть учтены системой как «хорошие» признаки в силу того, что они имеют низкую частоту в корпусе, на котором обучалась система. В обсуждаемом докладе авторы предложили использовать предварительную семантическую группировку лексики, полученную автоматической кластеризацией по большому неразмеченному корпусу текстов. Как отмечают авторы, при таком подходе для обучения системы нет необходимости в аннотированных корпусах большого объема. Система может обучаться и на небольшом множестве примеров для каждого подкласса объектов, размеченных вручную.

Другой способ «расширения» множества учитываемых признаков – это привлечение внешних баз знаний и онтологий. Использованию такого ресурса, как Википедия, в системах извлечения именованных сущностей обсуждался в целом ряде докладов: доклад группы исследователей из Японии [Higashinaka et al. 2012] и исследователей из Бирмингемского университета [Alotaibi, Lee 2012].

После того, как становится возможным выделять в тексте основных участников событий (персон, организаций), становится возможным извлекать информацию о самих событиях (задача извлечения фактов). В данном случае центральной является задача извлечения из текста предикатно-аргументной структуры, фреймов. Проблема в том, что в связных текстах достаточно часто все участники одного события не упоминаются в пределах одной предикации, а само событие упоминается в тексте более одного раза. При этом для его обозначения используются разные синтаксические конструкции. В одном предложении основной факт может быть выражен глаголом, в другом – отглагольным существительным. Задаче вычисления именных перифраз, обозначающих событие, выраженное глаголом в предыдущем контексте, был посвящен стендовый доклад японского коллектива [Tanaka et al. 2012]. Таким образом, собственно лингвистическая задача определения всех актантов глагола связана с задачей анализа связного дискурса и извлечения разных участников фактов из разных фрагментов текста.

6.2. Технологии извлечения информации с использованием знаний, извлекаемых из веба

Если извлечение информации из текста до некоторого момента было ограничено самим текстом, то в наше время такие системы могут использовать дополнительную информацию, доступную в Сети. Так, для того, чтобы решить, какая именно персона с имением John Smith упоминается в тексте, система может использовать дополнительные сведения обо всех Джонах Смитах, например из Википедии, и учесть область деятельности конкретного Джона Смита из текста. Привлечение открытых информационных ресурсов типа Википедии и других онтологий открытого типа не только позволяет разрешать так называемую онтологическую омонимию в тексте, но также используется для разрешения лексической омонимии, для извлечения тезаурусной информации (например, информации о таких отношениях, как гиперонимия, меронимия и т. п.). Данный подход нашел свое развитие в таком направлении автоматического анализа текста, как семантические технологии или технологии семантического веба. Эти технологии относительно новые, и им был посвящен учебный семинар «Использование источников веб-данных для углубленной автоматической обработки текста» («Exploiting web data sources for advanced NLP»), который был проведен исследователем университета Беркли Жераром де Мело.

Дело в том, что многочисленные исследования предыдущих лет показали, что достаточно точное извлечение информации из текста требует детального описания предметной области: онтологического моделирования. Представление об объектах, их атрибутах и связях позволяет разрешать языковую и онтологическую омонимию, извлекать более точно эти связи из неструктурированных источников. Онтологическое моделирование предметной области в рамках одного проекта – трудоемкая задача. При этом в Сети, как уже отмечалось, существует достаточно большое количество данных: онто-

логий, тезаурусов. В них в структурированном машиночитаемом виде представлена информация сразу о многих областях деятельности (например, в Википедии представлены «карточки» объектов). Кроме того, одно из направлений развития технологий семантического веба – это создание открытых связанных данных. В рамках этого направления доступные в Сети онтологии и тезаурусы, представленные в специальном машиночитаемом формате, связаны между собой. Это позволяет использовать структурированную информацию об одной и той же сущности, событии, понятии, представленную в разных источниках: в онтологиях высокого уровня (содержащих общие понятия, задающие базовое разбиение наблюдаемой действительности на категории), в детальных онтологиях для отдельных областей знаний (например, в биомедицине), в лингвистических ресурсах. Эта информация о свойствах и связях объектов помогает разрешать онтологическую омонимию (см. пример выше), помогает «достраивать» семантическую сеть, дополнять информацию о событиях, сущностях, абстрактных концептах, извлекаемых из коротких текстов, таких, например, как твиттер, сообщения в блогах, короткие новостные сообщения.

Использованию ресурсов семантического веба был посвящен ряд докладов на самой конференции. Как уже отмечалось в п. 7.1, энциклопедические ресурсы открытого типа, такие, как Wikipedia, DBPedia, Freebase и другие, используются в задачах распознавания именованных сущностей [Higashinaka et al. 2012; Alotaibi, Lee 2012]. Еще один доклад группы ученых из Индийского технологического института [Mukherjee, Bhattacharyya 2012], в котором рассматривалась технология использования данных Википедии и WordNet, был посвящен использованию данных ресурсов для категоризации видео, представленного на YouTube, по жанрам (комедии, фильмы ужасов, спорт и т. п.). Для категоризации обычно используют обучающие размеченные по жанрам / тематике тексты, представляющие собой названия фильмов, метаописания, комментарии пользователей. Эти тексты совсем небольшого объема, содержат мало слов, могут совсем не содержать лексемы, которые надежно указывают на определенную категорию. Это могут быть произвольные именованные сущности или лексемы, не встретившиеся в обучающем корпусе. В таком случае мы можем использовать сведения из Википедии, например, для того, чтобы приписать фильму, в название которого входит именованная сущность NBA (аббревиатура для Национальной ассоциации по баскетболу), категорию ‘спорт’. Для уточнения категории фильма по лексеме, не встретившейся в обучающей выборке, авторы предлагают использовать семантически связанные с этой лексемой слова из WordNet-a. Тогда для обучения системы достаточно для каждой категории иметь некоторое начальное множество слов WordNet-a. Такой подход позволяет использовать метод машинного обучения без учителя, то есть не требует первоначальной разметки обучающего корпуса экспертами. Применение аналогичных методов расширения контекста обсуждалось также в докладе исследователей из Миссурийского университета, посвященном анализу запросов на естественном языке [Roy, Zeng 2012].

6.3. Анализ социальных сетей

Нельзя оставить без внимания такую относительно новую область исследований в рамках автоматической обработки текста, как анализ текстов социальных сетей. С одной стороны, данное направление исследований чрезвычайно востребовано с точки зрения практического применения. Исследование поведения пользователей сети, их мнений, их связей с другими группами пользователей играет существенную роль при определении маркетинговых стратегий бизнеса, организации рекламы, исследовании общественного резонанса на те или иные события, вычислении интересов различных групп населения и т. д. Такой анализ позволяет различным интернет-сервисам распознавать потребности пользователя и адаптировать свою стратегию в соответствии с ними.

Если смотреть на тексты социальных сетей со стороны разработчиков различных систем автоматического анализа языка, то они представляют собой особый материал исследования. Как правило, это тексты небольшого или сверхкраткого размера (ср.,

например, микроблоги), при анализе которых невозможно опираться на частотную дистрибуцию языковых объектов в тексте. Кроме того, эти тексты представляют собой особый тип коммуникации, в котором много сокращений, искажений написания слов, нередко отсутствует синтаксическая правильность и связность. Таким образом, для многих задач неприменимы стандартные статистические методы АОТ, анализ этих текстов требует разработки новых более сложных математических моделей, привлечения более сложных лингвистических знаний.

На конференции было представлено достаточно большое количество докладов, которые были посвящены анализу такого рода текстов. Так, анализу микроблогов были посвящены такие доклады, как [Tanaka et al. 2012], упомянутый в п. 7.2 доклад [Roy, Zeng 2012], а также [Cassidy et al. 2012; Chen Y. et al. 2012; Duan et al. 2012] и др. Наибольшее внимание было уделено проблемам анализа текстов твиттера, обсуждались такие проблемы, как использование технологий семантического веба, данных открытых энциклопедий для выделения именованных сущностей в микроблогах, классификация блогов по темам, построение резюме на основе кластера твиттов и др.

7. АНАЛИЗ ТОНАЛЬНОСТИ И ИЗВЛЕЧЕНИЕ МНЕНИЙ

Анализ тональности текста, определение эмоциональной его окрашенности (subjectivity), определение позитивного или негативного отношения автора текста к оцениваемому или описываемому им объекту, событию (sentiment analysis), извлечение мнений (opinion mining) – все эти темы, связанные с оценочным компонентом элементов дискурса, пользуются большой популярностью в современных исследованиях. Они представляют собой активно развивающееся направление в области извлечения информации из текста. Неслучайно этому направлению был посвящен отдельный семинар. Определение положительной или отрицательной тональности текста, определение мнения относительно некоторого события или объекта действительности являются такими же чрезвычайно востребованными задачами, как и анализ социальных сетей. Они актуальны для совершенно разных областей – в сфере политики, при анализе масс-медиа и социальных сетей, в таких сферах бизнеса, как маркетинг, реклама, бизнес-разведка. В основе технологий автоматического анализа тональности лежат теоретические работы в области лингвистической интерпретации эмоций, теории субъективности, методы машинного обучения, в том числе различные классификационные методы (см. 8.1).

Основные решаемые задачи в этой области можно условно свести к следующим:

- теоретические подходы к описанию субъективной информации в тексте, уточнение теоретической психолингвистической базы оценки; разграничение эмоционально окрашенных (несущих субъективную оценку) и нейтральных текстов (фрагментов текстов);
- определение общей положительной vs. отрицательной оценки, содержащейся в тексте; определение оценки относительно некоторого оцениваемого объекта / аспекта оцениваемого объекта;
- создание специальных словарей экспрессивной и оценочной лексики.

7.1. Теоретические подходы к описанию субъективной информации в тексте. Определение общей положительной vs. отрицательной оценки, содержащейся в тексте

Без прояснения того, что же считать положительной или отрицательной оценкой, без рассмотрения ее психологических, социальных и других аспектов, без учета различных типов оценки (например, это может быть эмоциональная оценка или этическая и т. п.) трудно построить достаточно точную классификацию текстов и высказываний, содержащих оценку.

Вопросу о различных типах субъективной оценки в текстах, способах их выражения, методах их выявления был посвящен отдельный семинар «2nd Workshop on

sentiment analysis where AI meets psychology» («Второй семинар по анализу мнений: точки пересечения подходов искусственного интеллекта и психологии»). В рамках семинара рассматривались психологические и социологические аспекты теории оценки. Так, профессор Дж. Мартин из Университета Сиднея в своем докладе изложил основные положения теории оценки, а также рассмотрел, каким образом эта теория применяется к анализу дискурса с точки зрения функциональной грамматики М. Халлидея. Он рассмотрел три основных типа отношений, лежащих в основе оценочных суждений: эмоциональная оценка (affect), или эмоциональное состояние индивида, суждение (judgment), представляющее собой этическую или социальную оценку, и ценностная оценка (appreciation), подразумевающая эстетическую оценку какого-либо физического явления, а также способы их выражения в языке.

Для задачи определения текста как содержащего негативную или позитивную оценку на данный момент разработаны стандартные процедуры обучения – классификационные методы машинного обучения. Однако такие методы требуют уточнений. Для повышения качества данных методов требуются более глубокие и сложные подходы к проблеме, а также решение более «тонких» задач: не просто определение общей тональности текста, но извлечение оценки каких-либо конкретных параметров объекта и др.

Одним из краеугольных вопросов является вопрос о том, как провести границу между суждениями, в которых содержится какая-то оценка, и суждениями, в которых содержится объективная информация. Вопрос о субъективности в дискурсе оказывается важным не только в рамках теоретических подходов, но и в реальных практических системах. Вопрос о методах выявления оценочных суждений вызывает особый интерес в связи с повышенным интересом исследователей к автоматической обработке текстов социальных сетей. В этой связи следует упомянуть доклад группы исследователей из США [Biyani et al. 2012], в котором авторы ставят задачу определения субъективности в дискурсе на материале блогов. В качестве признаков, позволяющих выявить субъективную оценку, используются не только стандартные признаки – экспрессивно окрашенные слова, слова контекста и т. п., но и другие характеристики, такие, например, как структура форума, свойства диалоговых реплик. Как отмечают авторы, блог представляет собой диалог между участниками. Авторы различают субъективные и объективные блоги. При этом субъективные блоги, по гипотезе авторов, содержат больше реплик, больше участников обсуждения. Также с точки зрения свойств диалогических реплик в объективном блоге дискуссия разворачивается по относительно стандартному сценарию: вначале кто-то задает вопрос, потом на него отвечают, следующий вопрос инициатора обсуждения обычно относится к детализации и т. д. В субъективных блогах ответов на вопрос инициатора обсуждения, как правило, больше, поскольку они предполагают не просто сообщение некоторой объективной информации, но и выражение отношения к тому, что было сказано выше, со стороны разных участников. Кроме того, при высказывании своего отношения к некоторому факту участники нередко переходят к оценке личности отвечающего, таким образом, в ветке обсуждения образуются пустые сообщения. Таким образом, признаки, отражающие свойства и структуру диалогических реплик в блоге, помогают определить, является ли данное обсуждение оценочным или объективным.

7.2. Создание специальных списков / словарей оценочной лексики

Успешность применения метода машинного обучения в значительной степени может зависеть от тех признаков классифицируемых объектов, которые учитывает классификатор (см. п. 8.1). Для задачи извлечения мнений во многих системах такими признаками служит эмоционально окрашенная или оценочная лексика. Таким образом, одной из задач является составление специальных словарей оценочной лексики.

С одной стороны, в процессе развития систем анализа тональности были созданы специальные словари оценочной лексики. Например, для английского языка существу-

ет специальный тезаурус SentiNet[4]. С другой стороны, оценочная лексика, как правило, очень многозначна. Ее положительная или отрицательная окраска может зависеть от контекста, в частности от объекта оценки, от оцениваемого аспекта некоторого объекта (ср., например, *высокий результат в раскрываемости преступлений* vs. *высокий уровень преступности*). Поэтому вопрос о выделении специальных списков положительно и отрицательно окрашенных слов остается до сих пор актуальным. На конференции было представлено достаточно много докладов, посвященных этой теме (например, доклад российских исследователей [Chetviorkin, Loukachevitch 2012], а также [Biyani et al. 2012] и др.).

В докладе [Chetviorkin, Loukachevitch 2012] обсуждался метод построения словаря оценочных слов для оценки продуктов, не зависимого от конкретной предметной области. Авторы предложили методику построения словаря оценочных слов, независимых от конкретной предметной области, с использованием методов машинного обучения. В качестве корпусов для обучения использовались отзывы о четырех классах продуктов. В основу признаков, по которым лексика классифицировалась как оценочная или нет, легли различные квантитативные оценки специфичности слов-кандидатов, основанные на частоте встречаемости этих слов именно в отзывах по сравнению с текстами-описаниями в той же предметной области.

Предмет исследования авторов доклада [Biyani et al. 2012] (см. также п. 7.1) – оценочные слова с непостоянной (inconsistent) ориентацией оценки (как, например, *высокий*). Такие слова создают проблему при автоматическом извлечении оценочной лексики, поскольку для правильной их интерпретации необходимо учитывать контекстные слова, в частности, существительные, с которыми они сочетаются. Авторы предлагают подход, использующий машинное обучение, на основе размеченных по позитивно vs. негативно ориентированной оценке предложений. Предлагается учитывать информацию о синтаксических конструкциях в качестве признаков в системе машинного обучения для извлечения таких «амбивалентных» слов и правильного распознавания их оценочной интерпретации.

8. СОВРЕМЕННЫЕ ТЕХНОЛОГИИ РЕШЕНИЯ ЗАДАЧ В ОБЛАСТИ КОМПЬЮТЕРНОЙ ЛИНГВИСТИКИ

8.1. Машинное обучение

Как уже отмечалось выше, современные системы автоматической обработки текста, независимо от конкретных задач, будь это морфологический или синтаксический анализ или извлечение именованных сущностей (например, персон, организаций, географических названий), в значительной степени базируются на моделях машинного обучения. До сих мор методы машинного обучения являются «мейнстримом» автоматического анализа текста. Несмотря на то, что существует некоторый стандартный набор методов, применяемый в инженерной практике, эти методы не перестают совершенствоваться, возникают все новые, более сложные методы. На основе статистического анализа корпусов текстов больших объемов строятся различные математические модели, которые оказываются в состоянии приписывать правильные лингвистические характеристики элементам текста. Успешность многих методов определяется тем, какие лингвистические признаки будет учитывать система при обучении, какие метрики будут выбраны для их измерения. Вопросам о том, как должно быть устроено пространство признаков, как облегчить процедуру отбора релевантных лингвистических признаков, был посвящен целый ряд стендовых докладов. Например, в работе группы из Штутгартского университета [Heimerl et al. 2012] был рассмотрен специальный инструмент, который позволяет визуализовать статистические свойства признаков.

[4] http://sentiwordnet.isti.cnr.it

8.2. Методы краудсорсинга (Crowdsourcing)

Следует остановиться на еще одной современной интересной тенденции в развитии технологий обработки текста. Обсуждаемая ниже технология широко используется в различных областях, включая такую область, как моделирование знаний. Как известно, ранние эксперименты по машинному переводу столкнулись с такой проблемой: чтобы адекватно переводить на другой язык, необходимо учитывать реалии, о которых рассказывается в тексте, или, другими словами, модель предметной области. Именно поэтому от задачи анализа любых текстов исследователи в области моделирования понимания естественного языка (а также машинного перевода, экспертных систем) перешли к решению задач подобного рода на материале ограниченных предметных областей. Задача построить универсальную модель знаний, включающую все области человеческой деятельности, казалась невыполнимой.

Как уже отмечалось в п. 6.2, в последнее время в Сети появились многочисленные источники знаний о мире (такие как, например, Википедия), которые аккумулируют данные сразу обо всех областях деятельности. Они представляют собой знания многих людей, собранные в одном месте, такую коллективную модель мира. То, что когда-то казалось неподъемным для сил одного коллектива, стало возможным при распределенном создании различных информационных ресурсов многими людьми. Именно такой принцип был положен в основу активно развивающейся технологии сбора и обработки большого объема тех данных, которые нельзя доверить машине, – краудсорсинга.

Методы коллективного создания ресурсов, которые основываются на экспертных решениях и объем которых делает невозможным обработку данных силами одного коллектива в небольшие сроки, оказываются актуальными для различных разработок в области автоматического анализа текста. Помимо использования баз знаний, полученных методом краудсорсинга, развиваются также собственно лингвистические ресурсы, пополняемые самими пользователями по специально разработанным сценариям. Здесь актуальными оказываются следующие задачи: технология разбиения большой сложной задачи на множество мелких простых подзадач, с которыми может справиться малоквалифицированный пользователь; создание специальных инструментов, которые позволяют пользователю выполнять эти задачи (специального интерфейса для ответа на серию вопросов или интерфейса для разметки), разработка методов автоматизированной верификации данной разметки. В частности, применение технологии краудсорсинга использовалось в докладе группы ученых из Национального университета Сингапура, посвященном классификации временной референции [Ng, Kan 2012].

9. ИССЛЕДОВАНИЯ, НАХОДЯЩИЕСЯ НА СТЫКЕ ТЕОРЕТИЧЕСКОЙ, ЭКСПЕРИМЕНТАЛЬНОЙ И КОМПЬЮТЕРНОЙ ЛИНГВИСТИКИ

Развитие компьютерной лингвистики на данном этапе непосредственно связано не только с развитием технологий обработки текста, но и с повышенным интересом исследователей к задачам и достижениям теоретической и экспериментальной лингвистики.

Разработчики компьютерных интеллектуальных систем проявляют интерес к последним достижениям когнитивной науки, к исследованиям в области психо- и нейролингвистики. С одной стороны, последние достижения вычислительной лингвистики, когнитивной науки интегрируются в модели компьютерной обработки текста. С другой – модели, разрабатываемые в рамках обработки текста, применяются в различных «некомпьютерных» сферах, таких как изучение второго языка, исследования «удобочитаемости» текста (readability) и др.

9.1. Использование моделей вычислительной и когнитивной лингвистики в решении задач автоматической обработки текста

В связи с тем, что современные системы автоматического анализа текста включают такие глубинные уровни лингвистического анализа, как семантический и дискурсивный анализ текста, и при этом моделирование этих уровней стало возможно на принципиально новом технологическом уровне, возникает потребность и возможность интеграции достаточно сложных когнитивных и формальных вычислительных моделей в такие системы. В то же время в течение последних десятилетий в области теоретической лингвистики и когнитивной науки активно разрабатываются и совершенствуются модели организации семантической структуры языка, организации дискурса.

Исследования, посвященные интеграции таких моделей в системы автоматического семантического анализа, были представлены и на конференции COLING 2012. Так, в работе группы из Университета Амстердама [Le, Zuidema 2012] описывается метод автоматического семантического анализа, представляющий собой перевод предложений в формально-семантическое представление, основанное на семантике Монтегю [Montague 1970]. В задачи такого анализа входит вычисление значения предложения с использованием принципов формальной логики, на основе применения лямбда-исчисления. При данном подходе вычисление композиционального значения более крупных синтаксических единиц строится на основе значений единиц более низкого синтаксического уровня. Авторы доклада предлагают усовершенствовать метод семантической композиции для деревьев зависимости, заменив модель, основанную на лямбда-исчислении, графовой моделью. Такой подход, как показано в данной работе, позволяет применить обучение с использованием вероятностных моделей к семантическому анализу предложений.

В качестве примера использования моделей когнитивной психологии стоит упомянуть доклад [Roy, Zeng 2012] (см. также п. 6.2 и 6.3), посвященный семантическому представлению поисковых запросов. Как отмечается в докладе, успешность поиска определяется тем, насколько точно система может определить ключевые слова запроса. Один из методов определения заключается в сопоставлении запросу фрагмента семантической сети в виде графов, отражающих семантические связи между словами. В основе подхода авторов лежит трехмерная модель структуры интеллекта Джоя Пола Гилфорда [Guilford 1977], включающая три стороны: содержание, операции и результаты (интеллектуальные продукты). Авторы опираются на разработанное в рамках когнитивной психологии понятие семантической формы (представления). Предложенная авторами модель семантической сети включает такие компоненты, как элементы, классы, отношения, системы и т. п. Словам в запросе ставятся в соответствие фрагменты семантической сети. Далее в задачи системы входит установление некоторых «шаблонов» взаимодействия фрагментов семантических сетей, ассоциированных со словами в запросе. В своем исследовании авторы показывают, что обращение к моделям когнитивной психологии при решении задач автоматической обработки информации позволяет существенным образом улучшать результаты этой обработки. В то же время, эффективность применения таких моделей в практических системах может служить своего рода верификацией когнитивных моделей.

9.2. «Удобочитаемость» текста (readability)

Еще одно важное для современной компьютерной лингвистики направление, связанное с когнитивной наукой и психолингвистистикой, которое хотелось бы упомянуть, – это исследования свойств текста, определяющих его легкость/сложность для чтения. В рамках данного направления исследователи пытаются ответить на следующие вопросы: в какой степени некоторый текст «удобен» для чтения, какие характеристики влияют на степень его сложности. Вопросу о метриках удобочитаемости в различных языках, о методиках проверки качества текстов был посвящен целый ряд докладов, например,

[Hancke et al. 2012; Jameel et al. 2012] и др. В докладе коллектива из Тюбингенского университета [Hancke et al. 2012] обсуждалась проблема классификации текстов по степени сложности для чтения на немецком языке, рассматривались различные морфологические, лексические и синтаксические признаки. В работе исследователей из Гонконга [Jameel et al. 2012] предлагалась n-грамная модель (оценка цепочек из n слов) для оценки документов в ограниченных предметных областях по степени сложности для чтения (Ngram fragment sequence based unsupervised domain-specific document readability).

ЗАКЛЮЧЕНИЕ

В заключение хотелось бы обратить внимание на некоторые важные тенденции развития компьютерной лингвистики, нашедшие отражение и в докладах обсуждаемой в статье конференции.

Казалось бы, возможность применения различных методов машинного обучения привела к невостребованности экспертных лингвистических знаний: машина сама решает, какие лингвистические признаки значимы для решения тех или иных задач, сама извлекает правила, определяет, какие из них более важные. Однако, как видно из обсуждаемых выше задач, скорее можно констатировать вовлечение все более сложных уровней лингвистического анализа в сферу машинного обучения. При этом уровни сложности современных задач обработки текстов, таких, как извлечение временной референции в тексте, анализ структуры дискурса и многие другие, требуют экспертной разметки данных. То есть формальные теоретические модели не «уходят» со сцены, меняется их роль: они востребованы в разметке языковых данных для обучения. Иными словами, лингвистика «моделей» в разработках компьютерных систем постепенно преобразуется в лингвистику размеченных данных.

Другая значимая тенденция проявилась в особенностях формата конференции, который, как кажется, становится стандартом в современной компьютерной лингвистике. Это особый раздел работы, «Evaluation», посвященный количественному и качественному разбору результатов на фоне стереотипной матрицы результатов предшественников. Как известно, в последние десять лет активно проводились соревнования лингвистических компонентов, в которых ответы систем сравнивались с аннотированным «золотым стандартом» (CONLL, SemEval, EVALITA и др.). Материальным «выходом» этих соревнований стала публикация мини-корпусов, на которых проводились соревнования, их «золотой» аннотации и ответов участников. На нынешнем этапе развития исследований компьютерный лингвист чувствует себя обязанным протестировать свой алгоритм на всем пуле доступных результатов соревнований и оценить свое место в общем зачете. Заметим, что редко кому удается превзойти результаты предшественников по всем позициям, однако тщательный разбор отдельных удач и типичных неудач вносит маленький, но вклад в общий прогресс направления (как в смысле понимания преимуществ предложенного алгоритма, так и в смысле оценки структурных качеств коллекции, на которой проводилось обучение).

СПИСОК ЛИТЕРАТУРЫ

Agic 2012 – *Z. Agic*. K-best spanning tree dependency parsing with verb valency lexicon reranking pages // Proceedings of COLING 2012. URL: http://www.aclweb.org/anthology/C12-2001.

Alotaibi, Lee 2012 – *F. Alotaibi, M. Lee*. Mapping Arabic Wikipedia into the named entities taxonomy // Proceedings of COLING 2012: Posters. The COLING 2012 Organizing Committee. URL: http://www.aclweb.org/anthology/C12-2005.

Attia et al. 2012 – *M. Attia, P. Pecina, Y. Samih, Kh. Shaalan, J. van Genabith.* Improved spelling error detection and correction for Arabic // Proceedings of COLING 2012. URL: http://www.aclweb.org/anthology/C12-2011.

Atutxa et al. 2012 – *A. Atutxa, E. Agirre, K. Sarasola*. Contribution of complex lexical information to solve syntactic ambiguity in Basque // Proceedings of COLING 2012. URL: http://www.aclweb.org/anthology/C12-1007.

Avramidis 2012 – *E. Avramidis*. Comparative quality estimation: automatic sentence-level ranking of multiple machine translation outputs // Proceedings of COLING 2012. URL: http://www.aclweb.org/anthology/C12-1008.

Béchara et al. 2012 – *H. Béchara, R. Rubino, Y. He, Y. Ma, J. van Genabith*. An evaluation of statistical post-editing systems applied to RBMT and SMT systems // Proceedings of COLING 2012. URL: http://www.aclweb.org/anthology/C12-1014.

Bejček et al. 2012 – *E. Bejček, J. Panevová, J. Popelka, P. Straňák, M. Ševčíková, J. Štěpánek, Z. Žabokrtský*. Prague dependency treebank 2.5 – a revisited version of PDT 2.0 // Proceedings of COLING 2012. URL: http://www.aclweb.org/anthology/C12-1015.

Billingsley, Curran 2012 – *R. Billingsley, J. Curran*. Improvements to training an RNN parser pages // Proceedings of COLING 2012. URL: http://www.aclweb.org/anthology/C12-1018.

Biyani et al. 2012 – *P. Biyani, S.B.C. Caragea, P. Mitra*. Thread specific features are helpful for identifying subjectivity orientation of online forum threads // Proceedings of COLING 2012: Posters. The COLING 2012 Organizing Committee. URL: http://www.aclweb.org/anthology/C12-1019.

Botha et al. 2012 – *J.A. Botha, Ch. Dyer, Ph. Blunsom*. Language modelling of German compounds // Proceedings of COLING 2012. URL: http://www.aclweb.org/anthology/C12-1022.

Cassidy et al. 2012 – *T. Cassidy, H. Ji, L.-A. Ratinov, A. Zubiaga, H. Huang*. Analysis and enhancement of Wikification for microblogs with context expansion // Proceedings of COLING 2012. URL: http://www.aclweb.org/anthology/C12-1028.

Chen H.-B. et al. 2012 – *H.-B. Chen, H.-H. Huang, H.-H. Chen, Ch.-T. Tan.* A simplification-translation-restoration framework for cross-domain SMT applications // Proceedings of COLING 2012. URL: http://www.aclweb.org/anthology/C12-1034.

Chen Y. et al. 2012 – *Y. Chen, Zh. Li, L. Nie, X. Hu, X. Wang, T.-S. Chua, X. Zhang*. A semi-supervised Bayesian network model for microblog topic classification // Proceedings of COLING 2012. URL: http://www.aclweb.org/anthology/C12-1035.

Chetviorkin, Loukachevitch 2012 – *I. Chetviorkin, N. Loukachevitch*. Extraction of Russian sentiment lexicon for product meta-domain // Proceedings of COLING 2012. URL: http://www.aclweb.org/anthology/C12-1037.

Claveau 2012 – *V. Claveau*. Unsupervised and semi-supervised morphological analysis for information retrieval in the biomedical domain // Proceedings of COLING 2012. URL: http://www.aclweb.org/anthology/C12-1039.

Coyne et al. 2012 – *B. Coyne, A. Klapheke, M. Rouhizadeh, R. Sproat, D. Bauer.* Annotation tools and knowledge representation for a text-to-scene system // Proceedings of COLING 2012. URL: http://www.aclweb.org/anthology/C12-1042.

Delpech et al. 2012 – *E. Delpech, B. Daille, E. Morin, C. Lemaire*. Extraction of domain-specific bilingual lexicon from comparable corpora: compositional translation and ranking // Proceedings of COLING 2012. URL: http://www.aclweb.org/anthology/C12-1046.

Duan et al. 2012 – *Y. Duan, Zh. Chen, F. Wei, M. Zhou, H.-Y. Shum.* Twitter topic summarization by ranking tweets using social influence and content quality // Proceedings of COLING 2012. URL: http://www.aclweb.org/anthology/C12-1047.

Farkas, Bohnet 2012 – *R. Farkas, B. Bohnet*. Stacking of dependency and phrase structure parsers // Proceedings of COLING 2012. URL: http://www.aclweb.org/anthology/C12-1052.

Feng et al. 2012 – *M. Feng, W. Sun, H. Ney.* Semantic cohesion model for phrase-based SMT derivation // Proceedings of COLING 2012. URL: http://www.aclweb.org/anthology/C12-1053.

Fort et al. 2012 – *K. Fort, A. Nazarenko, S. Rosset*. Modeling the complexity of manual annotation tasks: a grid of analysis // Proceedings of COLING 2012. URL: http://www.aclweb.org/anthology/C12-1055.

Gareyshina et al. 2012 – *A. Gareyshina, M. Ionov, O. Lyashevskaya, D. Privoznov, E. Sokolova, S. Toldova.* RU-EVAL-2012: Evaluating dependency parsers for Russian // Proceedings of COLING 2012. URL: http://www.aclweb.org/anthology/C12-2035.

Goldberg, Nivre 2012 – *Y. Goldberg, J. Nivre*. A dynamic oracle for arc-eager dependency parsing // Proceedings of COLING 2012. URL: http://www.aclweb.org/anthology/C12-1059.

Gong et al. 2012 – *Zh. Gong, M. Zhang, Ch. Tan, G. Zhou*. Classifier-based tense model for SMT // Proceedings of COLING 2012. URL: http://aclweb.org/anthology/C/C12/C12-2041.pdf.

Guilford 1977 – *J.P. Guilford*. Way beyond the IQ. Creative education foundation. New York, 1977.

Hancke et al. 2012 – *J. Hancke, S. Vajjala, D. Meurers.* Readability classification for German using lexical, syntactic, and morphological features // Proceedings of COLING 2012. URL: http://www.aclweb.org/anthology/C12-1037.

Heimerl 2012 – *F. Heimerl, Ch. Jochim, S. Koch, Th. Ertl.* FeatureForge: A novel tool for visually supported feature engineering and corpus revision // Proceedings of COLING 2012. URL: http://www.aclweb.org/anthology/C12-2046.

Higashinaka et al. 2012 – *R. Higashinaka, K. Sadamitsu, K. Saito, T. Makino, Y. Matsuo*. Creating an extended named entity dictionary from Wikipedia // Proceedings of COLING 2012. http://www.aclweb.org/anthology/C12-1071.

Iida, Tokunaga 2012 – *R. Iida, T. Tokunaga.* A metric for evaluating discourse coherence based on coreference resolution // Proceedings of COLING 2012. URL: http://www.aclweb.org/anthology/C12-2048.

Inokuchi, Yamaoka 2012 – *A. Inokuchi, A. Yamaoka*. Mining rules for rewriting states in a transition-based dependency parser for English // Proceedings of COLING 2012. URL: http://www.aclweb.org/anthology/C12-1078.

Inoue et al. 2012 – *N. Inoue, E. Ovchinnikova, K. Inui, J. Hobbs*. Coreference resolution with ILP-based weighted abduction // Proceedings of COLING 2012. URL: http://www.aclweb.org/anthology/C12-1079.

Jameel et al. 2012 – *Sh. Jameel, X. Qian, W. Lam*. N-gram fragment sequence based unsupervised domain-specific document readability // Proceedings of COLING 2012. URL: http://www.aclweb.org/anthology/C12-1080.

Jones et al. 2012 – *B. Jones, J. Andreas, D. Bauer, K.M. Hermann, K. Knight*. Semantics-based machine translation with hyperedge replacement grammars // Proceedings of COLING 2012. URL: http://www.aclweb.org/anthology/C12-1083.

Kim et al. 2012 – *J. Kim, J. Nam, I. Gurevych*. Learning semantics with deep belief network for cross-language information retrieval // Proceedings of COLING 2012. URL: http://www.aclweb.org/anthology/C12-2057.

Kong, Zhou 2012 – *F. Kong, G. Zhou*. Exploring local and global semantic information for event pronoun resolution // Proceedings of COLING 2012. URL: http://www.aclweb.org/anthology/C12-1090.

Le, Zuidema 2012 – *Ph. Le, W. Zuidema*. Learning compositional semantics for open domain semantic parsing // Proceedings of COLING 2012. URL: http://www.aclweb.org/anthology/C12-1037.

Lesk 1986 – *M. Lesk*. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone // SIGDOC '86: Proceedings of the 5th annual international conference on systems documentation. New York, 1986.

Li et al. 2012 – *W. Li, J. Li, Y. Tian, Zh. Sui*. Fine-grained classification of named entities by fusing multi-features // Proceedings of COLING 2012. URL: http://www.aclweb.org/anthology/C12-2068.

Mathet et al. 2012 – *Y. Mathet, A. Widlöcher, K. Fort, C. François, O. Galibert, C. Grouin, J. Kahn, S. Rosset, P. Zweigenbaum*. Manual corpus annotation: giving meaning to the evaluation metrics // Proceedings of COLING 2012. URL: http://www.aclweb.org/anthology/C12-2079.

Meyer, Gurevych 2012 – *Ch.M. Meyer, I. Gurevych*. To exhibit is not to loiter: a multilingual, sense-disambiguated Wiktionary for measuring verb similarity // Proceedings of COLING 2012. URL: http://www.aclweb.org/anthology/C12-1108.

Mille et al. 2012 – *S. Mille, A. Burga, G. Ferraro, L. Wanner*. How does the granularity of an annotation scheme influence dependency parsing performance? // Proceedings of COLING 2012. URL: http://www.aclweb.org/anthology/C12-2082.

Miller et al. 2012 – *T. Miller, Ch. Biemann, T. Zesch, I. Gurevych*. Using distributional similarity for lexical expansion in knowledge-based word sense disambiguation // Proceedings of COLING 2012. URL: http://www.aclweb.org/anthology/C12-1109.

Mírovský et al. 2012 – *J. Mírovský, P. Jínová, L. Poláková.* Does tectogrammatics help the annotation of discourse? // Proceedings of COLING 2012: Posters. URL: http://www.aclweb.org/anthology/C12-2083.

Montague 1970 – *R. Montague*. Universal grammar // Theoria. 1970. 36 (3).

Mukherjee, Bhattacharyya 2012 – *S. Mukherjee, P. Bhattacharyya.* YouCat: Weakly supervised Youtube video categorization system from meta data & user comments using WordNet & Wikipedia // Proceedings of COLING 2012. URL: http://www.aclweb.org/anthology/C12-1114.

Nakazawa, Kurohashi 2012 – *T. Nakazawa, S. Kurohashi*. Alignment by bilingual generation and monolingual derivation. Eleftherios. Comparative quality estimation: Automatic sentence-level ranking of multiple machine translation outputs // Proceedings of COLING 2012. URL: http://www.aclweb.org/anthology/C12-1120.

Ng, Kan 2012 – *J.-P. Ng, M.-Y. Kan*. Improved temporal relation classification using dependency parses and selective crowdsourced annotations // Proceedings of COLING 2012. URL: http://www.aclweb.org/anthology/C12-1129.
Nguyen et al. 2012 – *L. Nguyen, M. Van Schijndel, W. Schuler*. Accurate unbounded dependency recovery using generalized categorial grammars // Proceedings of COLING 2012. URL: http://www.aclweb.org/anthology/C12-1130.
Roy, Zeng 2012 – *S.D. Roy, W. Zeng*. Computational cognitive model for semantic sub-network extraction from natural language queries // Proceedings of COLING 2012. URL: http://www.aclweb.org/anthology/C12-1045.
Rudnicka et al. 2012 – *E. Rudnicka, M. Maziarz, M. Piasecki, S. Szpakowicz*. A strategy of mapping Polish WordNet onto Princeton WordNet // Proceedings of COLING 2012. URL: http://www.aclweb.org/anthology/C12-2101.
Salloum, Elissa 2012 – *W. Salloum, N.H. Elissa*. A dialectal to standard Arabic machine translation system // Proceedings of COLING 2012. URL: http://aclweb.org/anthology/C/C12/C12-3048.pdf
Schwab et al. 2012 – *D. Schwab, J. Goulian, A. Tchechmedjiev, H. Blanchon*. Ant colony algorithm for the unsupervised word sense disambiguation of texts: Comparison and evaluation // Proceedings of COLING 2012. URL: http://aclweb.org/anthology/C/C12/C12-1146.pdf.
Shen et al. 2012 – *H. Shen, R. Bunescu, R. Mihalcea*. Sense and reference disambiguation in Wikipedia // Proceedings of COLING 2012. URL: http://www.aclweb.org/anthology/C12-2108
Tanaka et al. 2012 – *Sh. Tanaka, N. Okazaki, M. Ishizuka*. Acquiring and generalizing causal inference rules from deverbal noun constructions // Proceedings of COLING 2012: Posters. The COLING 2012 organizing committee. URL: http://www.aclweb.org/anthology/C12-2118.
Waszczuk 2012 – *J. Waszczuk*. Harnessing the CRF complexity with domain-specific constraints. The case of morphosyntactic tagging of a highly inflected language // Proceedings of COLING 2012. URL: http://www.aclweb.org/anthology/C12 -1170.
Xiao M. et al. 2012 – *M. Xiao, Y. Guo, A. Yates*. Semi-supervised representation learning for domain adaptation using dynamic dependency networks pages // Proceedings of COLING 2012. URL: http://www.aclweb.org/anthology/C12-1175.
Xiao X. et al. 2012 – *X. Xiao, D. Xiong, Y. Liu, Q. Liu, S. Lin*. Unsupevised discriminative induction of synchronous grammar for machine translation // Proceedings of COLING 2012. URL: http://www.aclweb.org/anthology/C12 -1176.
Zhai et al. 2012 – *F. Zhai, J. Zhang, Y. Zhou, Ch. Zong*. Machine translation by modeling predicate-argument structure transformation // Proceedings of COLING 2012. URL: http://www.aclweb.org/anthology/C12-1185.
Zhao, Marcus 2012 – *Q. Zhao, M. Marcus*. Long-tail distributions and unsupervised learning of morphology // Proceedings of COLING 2012. URL: http://www.aclweb.org/anthology/C12-1191.
Zhou et al. 2012 – *L. Zhou, W. Gao, B. Li, Zh. Wei, K.-F. Wong*. Cross-lingual identification of ambiguous discourse connectives for resource-poor language // Proceedings of COLING 2012. URL: http://www.aclweb.org/anthology/C12-2138.
Zhu et al. 2012 – *M. Zhu, J. Zhu, H. Wang*. Exploiting lexical dependencies from large-scale data for better shift-reduce constituency parsing // Proceedings of COLING 2012. URL: http://www.aclweb.org/anthology/C12-1194.

Сведения об авторах:

Светлана Юрьевна Толдова
Московский государственный университет им. М.В. Ломоносова
Национальный исследовательский университет «Высшая школа экономики»
toldova@yandex.ru

Ольга Николаевна Ляшевская
Национальный исследовательский университет «Высшая школа экономики»
Институт русского языка им. В.В. Виноградова РАН
olesar@gmail.com

Статья поступила в редакцию 12.02.2013.