Компьютерная лингвистика: развитие и прогноз практического применения технологий

А. С. Поречный, email: alex.porechny@mail.ru [1]
Е. В. Полицына, email: kathrin.beaver@mail.ru [1]
С. А. Полицын, email: pul_forever@mail.ru [1]

[1] Московский авиационный институт (национальный исследовательский университет)

Аннотация. *В статье выделяются периоды развития компьютерной лингвистики как прикладной дисциплины. На основе этого определяется возможное её положение в цикле зрелости технологий Gartner и делается прогноз перспективного развития компьютерной лингвистики.*

Ключевые слова: *компьютерная лингвистика, обработка естественного языка, цикл зрелости технологий, Gartner.*

Введение

Согласно закону, изложенному в диалектике Г. Гегеля, любое развитие происходит с помощью перехода количественных изменений в качественные. Такие переходы наблюдались в компьютерной лингвистике минимум дважды. Первый переход наблюдался, когда результаты множественных экспериментов и научных исследований перешли на качественно новый уровень, что позволило применить их в практических решениях, где автоматическая обработка языка является основной целью, например, в системах машинного перевода. Второй переход наблюдался, когда в результате научных и практических разработок накопилось множество доработок и улучшений, вследствие чего повысилась стабильность и качество результата работы автоматической обработке языка и снизились требования к ресурсам. Это позволило расширить практическое применение до областей, где ранее автоматическая обработка языка являлась важной, но вспомогательной или второстепенной функцией, например, поиск информации в социальных сетях, мессенджерах или Интернет-магазинах.

1. Развитие компьютерной лингвистики

Первым общеизвестным экспериментом автоматизации обработки естественного языка считается Джорджтаунский эксперимент 1954 года

по переводу предложений с русского языка на английский. Результат автоматического перевода более 60 предложений показал, что технически возможно автоматизировать обработку теста на естественном языке и произвести перевод. Несмотря на то, что перевод осуществлялся с существенными ограничениями, т.е. использовался словарь с всего 250 записями и 6 грамматическими правилами, начиная с этого момента в течение нескольких десятилетий происходит значительное увеличение числа новых экспериментов, разработки алгоритмов и подходов к машинному переводу, и в целом к обработке естественного языка [1].

За несколько десятилетий появляются системы по переводу текстов Systran, METEO, аналогичные системы разрабатываются в СССР для машин БЭСМ в ВИНИТИ РАН и в Математическом институте АН СССР. Ранее алгоритмы, используемые в системах тех лет, переводили текст буквально по словам, но постепенно добавлялось синтаксическое уподобление, т.е. порядок слов менялся в зависимости от требований, принятых в языке. При этом отмечалось, что для реализации полнотекстового перевода «… машины должны иметь большой объем «памяти», большое быстродействие…» [2].

Параллельно с машинным переводом разрабатываются системы по извлечению нужной информации из текста, в частности появляется система извлечения информации SMART, появляется крупномасштабная система извлечения информации Lockheed dialog System и т.д. Разрабатываются первые информационные системы по ранжированию поиска и индексируются массивы текстовых данных.

Вместе с этим активно развиваются системы и методы распознавания и синтеза речи, а также поддержания диалога. В области распознавания – это Audrey (распознавание произнесенных цифр), IBM Shoebox, алгоритм-DTW, устройство «Септрон». В области синтеза: Pattern playback, VODER, PAT/OVE, text-to-speech, Muse, методы LSD/LSE. В области поддержания диалога: ELIZA, Jabberwacky и др.

Примерно через 12-15 лет после Джордтаунского эксперимента наступает существенное разочарование в машинном переводе, постепенно это же происходит и с остальными задачами. Основной проблемой считалось, что «…недостаточное развитие теоретических исследований в области структурных и математических методов… тормозит практически важные работы по теории и практике машинного перевода, построению информационных языков и информационных машин, логической семантике и другим приложениям языкознания…» [4]

За это время формируется новая наука, которая появилась на стыке вычислительной техники и лингвистики – компьютерная лингвистика. На протяжении нескольких десятилетий постепенно очерчиваются и формируются цели и задачи, которые может решить и решает компьютерная лингвистика, различные исследователи по-разному выделяют задачи, но общий спектр остается неизменным:

– распознавание звучащей речи и синтез речи по тексту;
– распознавание входного текста;
– разработка системы «вопрос-ответ»;
– извлечение фактов и знаний на различных уровнях содержания информации в тексте, начиная с морфологического и закачивания семантическим и прагматическим;
– машинный перевод;
– и др.

Начиная с конца 80-ых годов прошлого столетия возвращается интерес к компьютерной лингвистике и ее задачам. Этому способствовало несколько факторов, например, бурное распространение компьютеров по всему миру привело с одной стороны к потребности иметь привычный интерфейс на естественном языке, а не ранний низкоуровневый интерфейс, состоящий из консоли и набора команд, подтверждением этому косвенно можно считать развитие графических интерфейсов в сторону взаимодействия с помощью компьютерной мыши. С другой стороны, распространение портативных и персональных ЭВМ и увеличение их мощности позволило разрабатывать более сложные алгоритмы анализа, а на фоне появления первых версий локальных и глобальных сетей возникают новые возможности по обработке текста для проведения поиска, объединения знаний и т.д, в целом возрастают потребности в решении задач компьютерной лингвистики.

За счёт этого массово появляются новые технологические компании, целью которых становится получение прибыли с помощью предоставления сервисов, основанных на достижениях компьютерной лингвистики и решении задач по автоматической обработке текста на естественном языке. Например, компания Google с ориентаций на поиск в сети Интернет с учетом «умной» обработки текста, позже Яндекс, Yahoo, Rambler, Mail и др. В области распознавания ABBYY, Google Voice и др. В области оценки тональности и эмоциональной окраски текста ВААЛ и др.

При этом продолжается формирование значимых концепций, например, теория «Смысл-Текст». Появляются словари, ориентированные на использование средствами вычислительной

техники, например, «Грамматический словарь русского языка», составленный А.А. Зализняком, который является основополагающим для словарей, активно применяемых в настоящее время для автоматической обработки естественного языка.

Можно считать, что в компьютерной лингвистике случился переход из количества в качество, когда количество разработанных алгоритмов и подходов стало достаточным, чтобы сделать рывок в качестве их работы, что позволяло решать уже практические задачи на приемлемом уровне.

При этом особенностью такого перехода в компьютерной лингвистике заключается в том, что алгоритмы решения задач в области исследуются и развиваются неравномерно, важными вехами были:

– Для машинного перевода – до нулевых годов XXI века системы ЭТАП, PROMT, ФРАП, RETRANS, Verbmobil, ПОЛИТЕКСТ, ДИАЛИНГ, Magic Gooddy и т.д.

– Для поиска и извлечения информации из текста – Knowledge, DARPA TIPSTER [3], АОТ и т.д.

– Для определения тональности текста – появление латентно-семантического анализа тональности в девяностых годах XX века, и бурное развитие методов на основе методов машинного бучения в средине нулевых.

– Для генерации текста – до девяностых рекуррентные нейросети, до 00-ых сети LSTM.

– Для распознавания речи – до девяностых Tangora от IBM, до нулевых VAL от BellSouth, EARS, GALE, до середины десятых годов Google Voice.

Таким образом, прослеживается неравномерность исследований в компьютерной лингвистике и ее применимости для решения различных востребованных на тот момент задач. Однако, решение такого рода задач компьютерной лингвистики в данных системах являлось целью создания таких систем, т.е. достижения в области компьютерной лингвистики являлись основной движущей силой в них.

2. Текущее положение и применение в различных компаниях

Начиная с середины нулевых годов XXI века, применение компьютерной лингвистики постепенно проникает в сферы, которые ранее не были связаны напрямую с обработкой текста и речи, в отличие от поисковых систем или машинного перевода.

Ярким примером того, как компьютерная лингвистика способствовала появлению новых технологий, является появление голосовых роботов-помощников, которые позволили разгружать операторов за счет предварительного опроса клиента, и чат-ботов,

которые выполняли схожие функции, но требующие только извлечения информации из текста.

Начиная с середины нулевых годов XXI века появляются первые решения в России (Медиалогия), которые направлены на осуществление оценки имиджа по информации и обсуждениям в СМИ, а начиная с середины десятых годов также начинается анализ социальных сетей.

В Интернет-магазинах, базах знаний (Wiki, Confluence и др.), мессенджерах (Телеграмм, WhatsApp и др.), социальных сетях появляется поиск информации по сообщениям с использованием средств различных уровней анализа, например, морфологического, что позволяет искать слова не только по точному совпадению и в т.ч. происходят попытки поиска информации по сообщениям с учетом смысла запроса. Принципиальное отличие от поисковых-систем (Яндекс, Google, Yahoo, Википедия и др.) заключается в то, что поиск идет по не общедоступному набору текстов, как следствие невозможно составить статистику и по ней построить релевантные ответы. Аналогичная проблема возникает с низкочастотными запросами поисковых-систем, например, по статистике Яндекса на 2016 год 100 млн. из 280 млн. запросов в день являются разовыми или уникальными запросами, не имеющую статистики. Для решения таких задач могут применяться искусственные нейронные сети, например, Яндекс использует свою сеть «Палех».

Начиная с десятых годов XXI века активно развиваются программы-ассистенты, у которых основной задачей является не распознавание речи или получение семантического запроса пользователя, а помощь пользователю с помощью алгоритмических подходов и машинного обучения вместо использования ресурсов оператора. Например, «Олег» от Тинькофф в финансовой сфере (2016 г.), «Алиса» от Яндекс (2018 г.), Siri от Apple (2011 г.), «Салют» от Сбер (2020 г.), Google Assistant (2016 г.) и т.д.

На текущий момент некоторые крупные компания, у которых основной продукт или сервис не связаны с автоматизаций обработки естественного языка, активно занимаются задачами распознавания речи, анализом тональности текста, поиском и извлечением информации из текста, развитием диалоговых систем и чат-ботов, синтезом речи и генерации текста. В частности, на рис.1 представлена диаграмма, показывающая развитие дополнительных компетенций в разных российских компаниях [5].

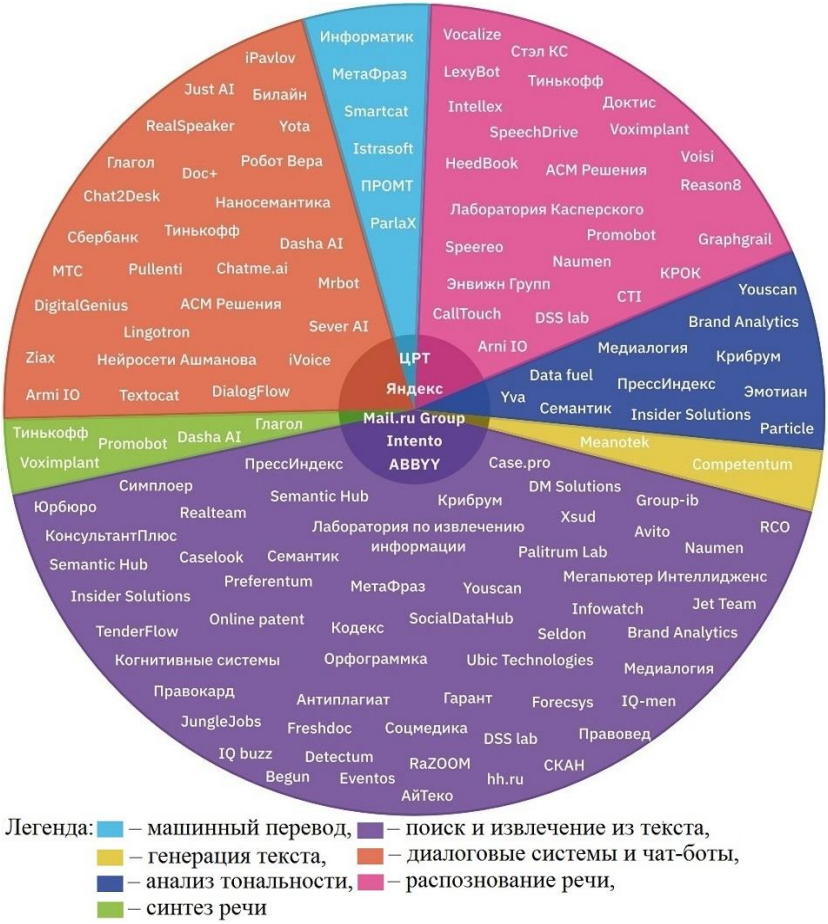

Рис. 1. Диаграмма областей автоматической обработки языка

Таким образом, постепенно появляются системы, в которых применение технологий компьютерной лингвистики не является необходимым для решения основных задач, но повышает удобство и эффективность использования таких систем.

3. Положение компьютерной лингвистики в цикле зрелости технологий Gartner

Безусловно, компьютерная лингвистика не была единственным направлением, которое развивалось с появлением ЭВМ. Также развивалось 3D-моделирование, облачные технологии и многие другие.

В процессе анализа различных технологий компания-исследователь Gartner приходит к понимаю существования условного цикла зрелости технологий, через который проходят все новые технологии и технологические компании [6].

На рисунке 2 изображен «цикл зрелости технологий Gartner». Цикл состоит из резкого и бурного «запуска технологии», который достаточно быстро достигает «пика завышенных ожиданий», далее происходит резкое разочарование в технологии, т.к. она не позволяет быстро решить задачу как ожидалось и как следствие технология попадает на этап «пропасти разочарования». Далее, если к технологии полностью не угас интерес, наступает «склон просвещения», который заключается в постепенном и поступательном развитии технологии, который перетекает в «плато продуктивности», когда технология становится достаточно удобной и широко применяется [6].

На основании исследований Gartner на кривой зрелости изображаются технологии и направления с точки зрения практического использования, поэтому компьютерная лингвистика представлена рядом отдельных направлений, связанных с автоматической обработкой естественного языка. Невозможность полной формализации естественного языка приводит к тому, что решения для каждой задачи развиваются отдельно от остальных, а прогресс по одной задаче не всегда влияет на прогресс другой. Например, достижения в области машинного перевода не влияют напрямую на синтез или распознавание речи.

Очевидно, что машинный перевод имел «пик завышенных ожиданий» в первые 12-15 лет после Джорджтаунского эксперимента, далее наступила «пропасть разочарования» и постепенно поступательно происходит развитие, которое на данный момент можно считать успешным, т.к. последние годы качество машинного перевода для большинства языков стало выше. В то же время «пик завышенных ожиданий» диалоговых систем и чат-ботов наступил значительно позже, и они до сих пор находятся в начале «склона просвещения» и т.д.

Для более корректной оценки положения в цикле зрелости компьютерной лингвистики необходимо учитывать наличие различных задач, которые исследуются неравномерно, а также учитывать, как

минимум два перехода в качестве результатов компьютерной лингвистики.

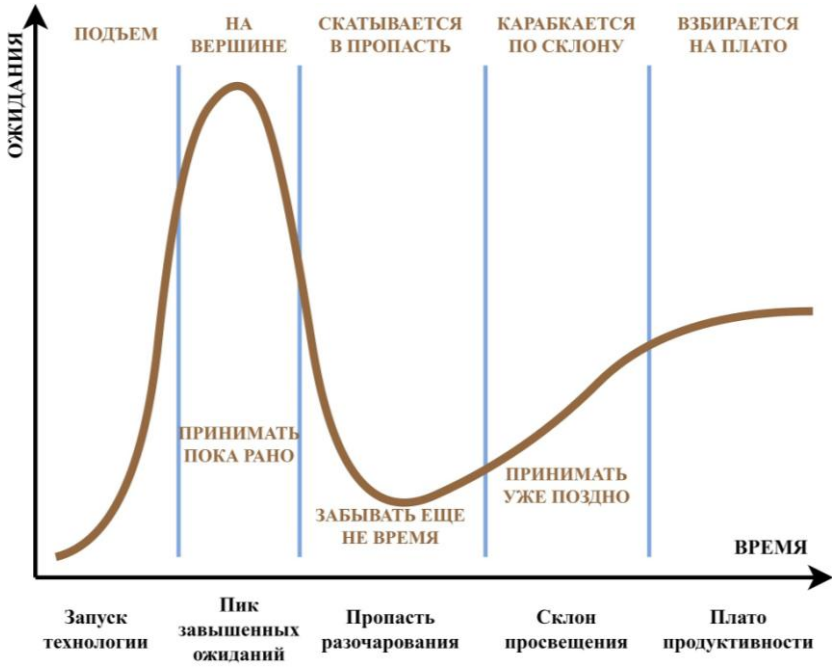

Рис. 2. Цикл зрелости технологий Gartner

4. Прогноз практического применения достижений компьютерной лингвистики

Учитывая график Gartner, можно предположить, что для некоторых задач уже произошло несколько переходов от количества к качеству, которые позволяют преодолеть «склон просвещения», например, сложно оспорить утверждение, что машинный перевод значительно повысил результаты несмотря на то, что экспертный перевод пока что остается более качественным.

Другие задачи, которые находятся на «склоне просвещения» или уже проходят его, все еще не применяются широко ввиду того, что не удалось еще определить удобный и простой способ использовать наработки компьютерной лингвистики для решения вспомогательных практических задач. Тем не менее, некоторые попытки создать библиотеки, которые помогают разрабатывать программы с использованием обработки естественного языка разрабатывались,

например, для машинного обучения это TensorFlow, Theano, Keras, scikit-learn, Lasagne, Caffe, DSSTNE, Wolfram Mathematica, на основе алгоритмических подходов библиотеки TAWT [7], GATE, LingPipe, UIMA, Texterra и т.д.

При этом существуют задачи компьютерной лингвистики, которые относительно недавно прошли «пропасть разочарования». Например, диалоговые технологии или чат-боты, которые дают релевантные ответы на запросы на естественном языке.

Заключение

Компьютерная лингвистика развивается уже несколько десятков лет, а параллельно с этим появляется все больше задач, где необходимо применение автоматической обработки языка. Причем такие задачи появляются в различных промышленных системах, начиная с тех, в которых основной целью является обработка языка, например, системы машинного перевода, поисковые системы и т.д., заканчивая теми, где обработка языка является дополнительной или вспомогательной задачей, например, задача морфологического поиска в базах знаний или в мессенджерах, социальных сетях и т.д.

Тем не менее в отличие от некоторых высокотехнологических компаний, которые имеют свои разработки в области автоматической обработки языка, технологии автоматической обработки языка остаются все еще недостаточно доступными, чтобы применять их повсеместно. В основном это связанно с тем, что существующие программные решения обработки языка требуют значительных знаний в области компьютерной лингвистики.

Технологическое развитие задач идет неравномерно и некоторые из них находятся в «пропасти разочарования» или на ранних стадиях «склона просвещения», но некоторые – на «склоне просвещения» и на ранних стадиях «плато продуктивности».

Учитывая то, что развитие происходит за счёт перехода количества в качество, можно сделать вывод, что в настоящее время для некоторых задач компьютерной лингвистики накапливается достаточно программных решений и вместе с ним компетенций, чтобы такие решения с одной стороны, повысили качество своей работы, с другой стороны стали более доступными, а их программные интерфейсы более интуитивно понятными.

Список литературы

1. Reynolds, A. Craig The conference on mechanical translation. // Mechanical Translation: 1954. – С. 47-55.

2. Ляпунов, А.А. Использование вычислительных машин для перевода с одного языка на другой. / А.А. Ляпунов, О.С. Кулягина // Природа. – Академия наук СССР. – 1955г. – №8 (август). – С. 83-95.

3. Kaufmann, M. Tipster Text Program Phase III: Proceedings of a Workshop held at Baltimore // MD, USA, Kaufmann, October 13-15, 1998. – pp. 39-40. ISBN 978-1-55860-610-4.

4. Григорьев, В.И. О развитии структурных и математических методов исследования языка / В.И. Григорьев // Вопросы языкознания. – 1960г., – № 4. – С. 153-155.

5. Альманах. Искусственный интеллект. Обработка естественного языка, распознавание и синтез речи // Центр Национальной технологической инициативы на базе МФТИ по направлению «Искусственный интеллект». – 2019г. – №2 (сентябрь). – 103с.

6. Цикл зрелости технологий Gartner [Электронный ресурс]: статья аналитического агентства – Режим доступа – https://www.tadviser.ru/index.php/Статья:Gartner_Hype_Cycle_for_Emerging_Technologies_(Цикл_зрелости_технологий_Gartner)

7. Полицына, Е.В., Алгоритмы автоматизации анализа текста на русском языке для решения прикладных задач с применением фреймворка TAWT / Е.В. Полицына, С.А. Полицын, А.С. Поречный // Программные продукты и системы. 2021. Т. 34. № 2. С. 257–268. DOI: 10.15827/0236-235X.134.