ПРАКТИЧЕСКИЕ РЕЗУЛЬТАТЫ РОССИЙСКИХ ИССЛЕДОВАТЕЛЬСКИХ ГРУПП В СФЕРЕ ОБРАБОТКИ ЕСТЕСТВЕННОГО ЯЗЫКА, ПОЛУЧИВШИЕ ПРАВОВУЮ ОХРАНУ (2013-2017)

Л.Р. Комалова
Институт научной информации по общественным наукам Российской академии наук,
Московский государственный лингвистический университет
Москва

Роберт Дейл [1] описывает категории продуктов на основе обработки естественного языка, которые в настоящее время встроены в реальные секторы мировой экономики и являются наиболее перспективными в отношении их коммерциализации. Отмечается, что именно в последние несколько лет коммерциализация этой сферы исследований стала значительно более ощутимой, несмотря на то, что первые шаги были предприняты еще 35 лет назад.

Со ссылкой на исследование Т. Джонсона [2], в котором определены основные категории рынка технологий на основе обработки естественного языка (см. табл. 1), автор отмечает, что в настоящее время первая, вторая и третья категории трансформировались в единую категорию «диалоговые системы» (Conversational Systems); четвертая расширилась до «анализа текста» (Text Analytics); пятая реализуется в средствах проверки грамматики (Grammar Checking); шестая осталась в неизменном виде; седьмая стала дополнением продуктов, предназначенных для распознавания речи, встроенных в ведущие операционные системы.

В качестве подтверждения того, что данные категоризации продуктов на основе лингвистических технологий применимы и на российском рынке искусственного интеллекта, можно привести интервью с руководителем компании «Endurance» [3] и программу Международной конференции о применении искусственного интеллекта и анализа данных 2018 года [4].

Таблица 1. Категории интеллектуальных продуктов на базе технологий обработки естественного языка

№	Категория	Определение
1.	Интерфейсы баз данных: мейнфреймы и компьютеры для решения исследовательских задач (Database interfaces: mainframes and minis)	Портативные или интегрированные интерфейсы для СУБД и др. приложений
2.	Интерфейсы баз данных: микропроцессоры (Database interfaces: micros)	ПК или подобные интерфейсы к программному обеспечению или онлайновым службам
3.	Диалоговые интерфейсы (Dialogue interfaces)	Диалоговые интерфейсы для таких комплексных систем как экспертные системы или приложения типа ICA1
4.	Контекстное сканирование (Content scanning)	Обработка сообщений или других полуформатированных текстов и принятие решения о действии или маршрутизации
5.	Редактирование текста (Text Editing)	Проверка стиля и грамматики текста, формулировка предложений по корректировке текста
6.	Машинный перевод (Machine Translation)	Компьютерный перевод текстов с исходного естественного языка на целевой естественный язык
7.	Машинопись (Talkwriters)	Транскрипция устной речи в орфографической форме

Примечание: по материалам [1, p. 642].

Обзор результатов интеллектуальной деятельности научных коллективов, работающих в российских организациях, проводится на основе анализа патентной документации на изобретения. Принимая во внимание всю трудоемкость и затратность процедуры подачи заявки и регистрации изобретения (по сравнению с регистрацией, например, авторского права на базу данных или программу для ЭВМ), предполагается, что разработчики, решившие получить «охранную грамоту» на свою разработку, провели анализ рынка, выявили аналоги и определили конкурентоспособность своего продукта или технологии. В связи с этим для решения исследовательских задач настоящей статьи представляется оправданным обращение именно к патентным документам, а не анализ научных статей, описывающих способы обработки естественного языка.

Целью проводимого анализа патентной документации было выявление соответствия патентуемых продуктов и технологий современным трендам в области обработки естественного языка, зафиксированным в работе [1], и определение наиболее ожидаемых (с позиций российской науки) областей коммерциализации результатов интеллектуальной деятельности в области обработки естественного языка. Патентный поиск

проводился по документам информационной системы Роспатента, локализованной на сайте Федерального института промышленной собственности [5]. Учитывались следующие параметры поиска:

– поиск проводился по базе данных патентных документов РФ на русском языке;
– поиск осуществлялся по рефератам и текстам заявок на государственную регистрацию изобретений;
– основная область запроса формулировалась следующим образом «язык* OR реч* OR дискурс* OR текст*»;
– глубина поиска составила пять лет (2013-2017г.);
– рассматривались документы со статусом «действует».

В результате поиска по заданным параметрам было обнаружено 549 патентов[1]. Из выборки были исключены изобретения, правообладателями которых являются представители других государств (Китай (13 патентов), Япония (5), США (5), Корея (4), Испания (2), Германия (2), Нидерланды (1), Канада (1)), а также изобретения, в описаниях которых ключевые слова не соответствовали исследуемой области (например, патенты на изобретения в области определения языковой компетенции, дислексии, социального поведения; в области медицинской коррекции физиологической целостности языка и желудочно-кишечного тракта; в области пожаротушения (языки пламени), кулинарии (язык как мясное изделие), технических устройств, не относящихся к обработке естественного языка). В итоге исследуемую выборку составили 76 патентов.

Отобранные для анализа патенты можно условно разделить на две категории: 1) обработка звучащей речи и 2) обработка письменного текста. Динамика регистрации изобретений на территории РФ по исследуемому предмету в соответствии с датой регистрации права и правообладателем отражена на рис. 1–3.

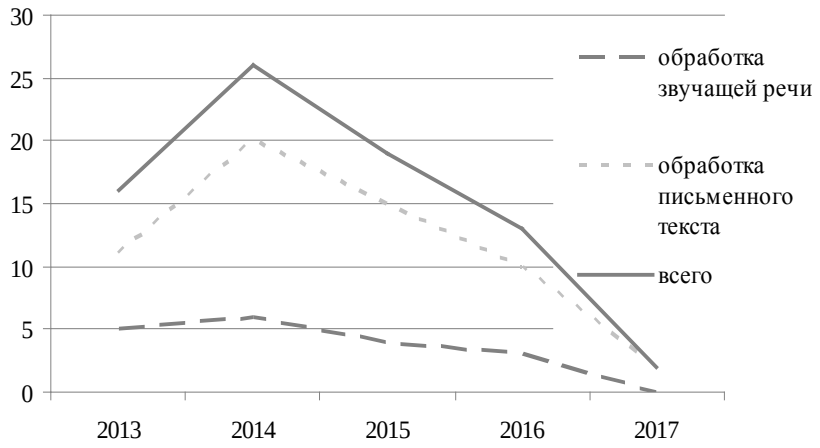

Рис. 1. Распределение результатов поиска по годам

Наибольшее количество зарегистрированных изобретений из анализируемой выборки приходится на 2014 год. Эта тенденция наблюдается как для способов и устройств для обработки звучащей речи, так и для обработки письменного текста. Следует отметить, что количество изобретений для обработки письменного текста значительно превышает количество изобретений для обработки звучащей речи.

Основными «игроками» на рынке изобретений по предмету поиска являются коммерческие организации (67% исследуемой выборки), а основным «поставщиком» выступает группа компаний «ABBYY» (за 2013–2017г. зарегистрировано 36 изобретений, что составляет 47% исследуемой выборки), компания «Яндекс» (6 изобретений, 8%) и компания «Лаборатория Касперского» (4 изобретения, 5%). Наряду с коммерческими организациями разработку способов и устройств обработки естественного языка осуществляют вузы (13%) и частные лица (16%). Четыре патента зарегистрировано «Государственным научно-исследовательским испытательным институтом проблем технической защиты информации Федеральной службы по техническому и экспортному контролю» (5%), два — «Краснодарским высшим военным училищем им. генерала армии С.М. Штеменко» (3%). Было найдено два патента, патентообладателями которых являются научные учреждения в системе РАН (3%), и один патент, правами на которое совместно распоряжаются организация (вуз) и частное лицо (изобретатель).

[1] Для сравнения: по тем же параметрам поиска в базе ФИПС «Программы для ЭВМ, БД и ТИМС» было найдено 6079 авторских свидетельств на программы для ЭВМ, 1144 авторских свидетельства на базы данных.

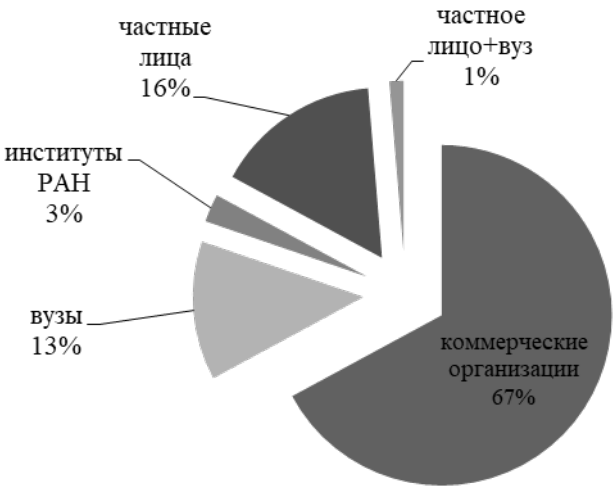

Рис. 2. Распределение результатов поиска по типам патентообладателей

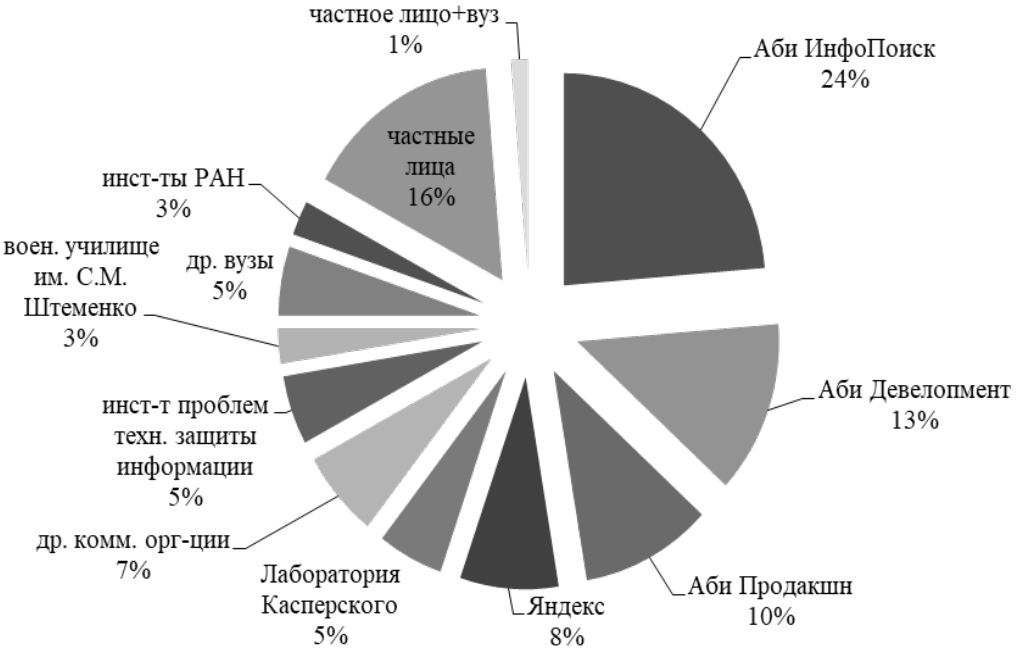

Рис. 3. Распределение результатов поиска по патентообладателям

Качественный анализ содержания патентов, релевантных условиям поискового запроса, позволил выделить следующие тенденции. Российские изобретения в области обработки естественного языка, зарегистрированные на территории РФ, можно разделить на следующие тематические группы[2]:

- анализ текста (информационный поиск, классификация, реферирование, определение языка текста, фильтрация и т.д. — 35 патентов);
- декодирование звучащей речи (10 патентов);
- преобразование изображения, содержащего текст, в машиночитаемый формат (8 патентов);
- перевод письменного текста с языка оригинала на язык запроса (4 патента);
- анализ речи в зашумленных условиях (4 патента);
- средства безопасности передачи текстовых данных (3 патента);
- разметка письменного текста (3 патента);
- синтез письменного текста (2 патента);
- редактирование текста (2 патента);

[2] Группировка материала проводилась исходя из анализа текстов реферата и (если реферата было недостаточно) на основе описания изобретения в идентифицирующих материалах к патенту.

– синтез письменного текста на основе устного сообщения (1 патент);
– создание текстовых корпусов (1 патент);
– синтез звучащей речи на основе письменного текста (1 патент);
– перевод устного сообщения с языка оригинала на язык запроса (1 патент);
– средства безопасности речевых данных (1 патент).

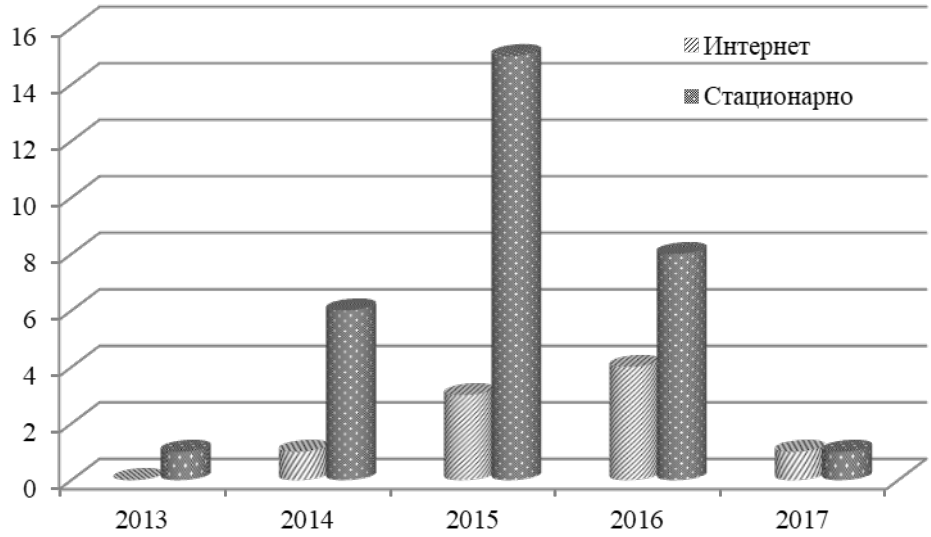

Рис. 4. Распределение результатов поиска по типу реализации изобретения

В анализируемой выборке четко выделяется два типа разработок: изобретения, работающие стационарно, и изобретения, связанные с Интернетом. Изобретения второго типа в исследуемой выборке представлены меньшим числом (рис. 4) патентов и направлены на решение следующих задач:
– повышение эффективности использования резерва производительности антивирусного сервера, фильтрации спама;
– повышение безопасности передачи данных по электронной почте, целостности электронных документов, надежности обнаружения фишинга;
– повышение эффективности использования приложений для составления списков;
– повышение эффективности взаимодействия пользователя с чатовой информационной системой;
– расширение средств анализа цифрового изображения;
– повышение эффективности создания краткого изложения (реферата) цифрового текста, обработки поискового запроса;
– повышение эффективности переводческого сервиса на базе электронного сообщества, скорости перевода с языка оригинала на язык запроса;
– повышение эффективности сбора целевой информации в чате;
– повышение точности определения пользовательского текста в текстовом контексте при автоматическом формировании корпуса текста по заданным параметрам;
– *обработка звучащей речи на естественном языке*;
– предоставление пользователю доступа к дистанционным услугам;
– прогнозирование эффективности речевого воздействия дискурса на иностранном языке.

В результате проведенного анализа можно сделать вывод о том, что актуальные (2013–2017 гг.) российские разработки в области обработки естественного языка в целом соответствуют современным трендам, зафиксированным в работе Р. Дейла. Наиболее представленными в коммерческом плане российскими изобретениями являются разработки в области анализа письменного текста (Text Analytics), в незначительной степени представлены разработки в области машинного перевода (Machine Translation) и продукты по распознаванию звучащей речи. Перспективной для отечественных исследователей на внутреннем рынке является сфера разработок в области диалоговых системы (Conversational Systems).

Исследование выполнено в рамках гранта №18-18-00477 Российского научного фонда (РНФ).

ЛИТЕРАТУРА

1. Dale R. Industry watch: the commercial NLP landscape in 2017 // Natural Language Engineering. 2017. Vol. 23. № 4. Pp. 641–647. DOI: 10.1017/S1351324917000237.
2. Johnson T. Natural language computing: the commercial applications // The Knowledge Engineering Review. 1984. Vol. 1. № 3. Pp. 11–23. URL: https://doi.org/10.1017/S0269888900000588 (дата обращения: 15.01.2018).
3. Создание чат-ботов: Наукоемкая разработка или возможность заработать на AI? URL: https://4science.ru/articles/Georgii-Fomichev-iz-Endurance-vistupit-na-konferencii-AI-Conference-2018-po-teme-chat-botov (дата обращения: 05.02.2018).
4. Международная конференция о применении искусственного интеллекта и анализа данных. URL: https://aiconference.ru/ru (дата обращения: 31.01.2018).
5. ФИПС. Федеральный институт промышленной собственности. URL: http://www1.fips.ru (дата обращения: 09.02.2018).