Санкт-Петербургский государственный университет

ТАРАСОВА Маргарита Алексеевна

Выпускная квалификационная работа

Моделирование вопросов в диалоге (на материалах интервью в региональной прессе)

Уровень образования: магистратура

Направление 45.04.02 «Лингвистика»

Основная образовательная программа ВМ.5805 «Компьютерная и прикладная лингвистика»
Профиль «Компьютерная лингвистика»

Научный руководитель:
доцент, Кафедра математической лингвистики,
Николаев Илья Сергеевич
Рецензент:
ст. науч. сотрудник,
НИУ ВШЭ
в Санкт-Петербурге,
Паничева Полина Вадимовна

Санкт-Петербург
2021

Содержание

Введение

Одним из направлений в области искусственного интеллекта сегодня является обработка текстов на естественном языке. На анализе естественного языка основана разработка вопросно-ответных систем (от англ. Question Answering Systems) – информационных систем, способных принимать вопросы и отвечать на них. В отличие от поиска по ключевым словам, вопросно-ответная система предполагает краткий и «осмысленный» ответ, сформированный системой в результате анализа источников некоторой базы знаний.

Точность и полнота ответа в первую очередь зависит от анализа семантики вопроса. Как показал обзор исследований (Воробьёв 2020, Соловьёв 2010, Сулейманов 2016), модуль обработки вопроса пользователя считается самым ответственным и трудоёмким этапом в работе вопросно-ответной системы. Для этой задачи применяются различные методы семантического анализа: присвоение семантических тэгов заранее заданной таксономии, использование морфологических шаблонов, построение синтаксических деревьев, использование статистических методов, но все они используются, как правило, для английского языка.

В.П. Захаров относительно русского языка в контексте решения задач компьютерной лингвистики точно заметил: «Русский язык относится к категории самых сложных языков для семантического анализа, поэтому любые задачи, связанные с автоматической обработкой русскоязычного текста приходится рассматривать через призму языковой сложности» (Захаров 2015: 34). Незакреплённый порядок слов, богатство лексики, вариативность выбора языковых средств для достижения коммуникативной цели, смысловая неполнота (эллипсис), разрешение неоднозначности – и это только часть проблем, с которыми сталкиваются компьютерные лингвисты, создающие вопросно-ответные системы на русском языке. Тем не менее, вопросно-ответные системы постепенно разрабатываются и начинают применяться в России: в сфере бизнеса, экономики, образования, но пока не нашли

применения в издательском деле. Именно поэтому нам представляется важным описать пути создания и способ работы вопросно-ответной системы для регионального интернет-издания.

Объектом нашего исследования являются вопросы в интервью регионального интернет-издания, а **предметом** – способы распознавания их семантики для построения вопросно-ответной системы.

Цель данной выпускной квалификационной работы – моделирование русскоязычной вопросно-ответной системы для электронного республиканского издания ЯСИА.

В качестве **материала исследования** мы взяли вопросы и ответы интервью из региональной электронной газеты **ЯСИА** ***(ysia.ru)*** – Якутское-Саха информационное агентство – объёмом более 620.000 токенов.

Информационно-издательский холдинг «Сахамедиа», помимо ЯСИА, осуществляет выпуск печатных газет «Якутия» (на рус. яз) и «Саха Сирэ» (на якут. яз.). Учредителем и читателем ЯСИА является правительство республики. Это важный экстралингвистический фактор, который свидетельствует о том, что материал газет и электронного издания публикуется на русском/якутском литературном языке, чем обусловлен выбор данного издания в качестве материала исследования.

В первую очередь интернет-газета освещает деятельность правительства республики и изменения в законодательстве, рассказывает о жизни местного населения, а также анонсирует мероприятия, предупреждает о погодных изменениях. Жанровая специфика электронного издания достаточно широкая: это информационные жанры, аналитические и художественные.

Для достижения поставленной цели необходимо решить следующие **задачи:**

1) выявить языковые особенности интервью (источника информации для вопросно-ответной системы) и описать основные классификации вопросов в лингвистике и журналистике;

2) изучить, как метод моделирования применяется сегодня для решения задач в области прикладной лингвистики;
3) рассмотреть современные подходы построения вопросно-ответных систем;
4) собрать и выполнить лингвистическую обработку корпуса вопросов и ответов, включающую токенизацию, лемматизацию, проверку стоп-словарём и приведение к словарной форме всех слов;
5) обучить Word2vec модель корпуса;
6) протестировать пользовательские вопросы в созданной системе;
7) на ограниченной выборке пользовательских вопросов оценить точность, полноту и F-меру.

Методы и приемы исследования. В данной работе применяются методы дистрибутивной семантики и корпусной лингвистики. Кроме того, используются следующие библиотеки:

1. beautifulsoup4 для сбора вопросов и ответов;
2. re для написания токенизатора;
3. gensim для построения Word2vec модели корпуса;
4. numpy для преобразования пользовательских вопросов в векторы.

Теоретико-методологическую базу диссертации составили работы следующих учёных, специалистов в области компьютерной и прикладной лингвистики, информационного поиска, стилистики текста и журналистики: Ю.Д. Апресяна, К.Е. Белоусова, Е.М. Валгиной, С.П. Воробьёва, В.П. Захарова, И.В. Ивановой, О.С. Кожуновой, И.Ю. Морозова, В.А. Мочалова и А.В. Мочаловой, Г.Я. Солганик.

Научная новизна исследования заключается в том, что в данной работе впервые была сделана попытка создания русскоязычной вопросно-ответной системы, основанной на модели дистрибутивной семантики Word2vec, для электронного регионального издания.

Гипотеза: Word2vec позволяет учитывать семантические особенности текстов вопросов и ответов, поэтому эта модель может применяться для построения вопросно-ответных систем.

Теоретическая значимость работы состоит в разработке методики построения вопросно-ответной системы для регионального интернет-издания.

Практическая значимость работы заключается в том, что результаты исследования могут быть использованы для развития интернет-издательств и оптимизации вычленения информации, необходимой для пользователя.

Структура работы. Работа состоит из двух глав, введения, заключения, списка использованной литературы и приложения. В первой главе нами описаны теоретические предпосылки исследования: описаны языковые особенности текста интервью, рассмотрены основные классификации вопросов, рассмотрены подходы к разработке вопросно-ответных систем. Вторая глава посвящена экспериментам с корпусом вопросов и ответов региональных интервью: выявление специфики регионального текста (базы знаний), описание процедуры сбора вопросно-ответных данных и их обработка, построение Word2vec модели, тестирование и оценка результатов. Общий объём работы – 80 стр., основное содержание изложено на 72 стр. В работе содержится 4 таблицы и 9 рисунков.

Глава 1. Моделирование вопросно-ответной системы на основе текстов интервью

1.1. Языковые особенности интервью

Как показал анализ литературных источников, интервью в региональной прессе изучаются крайне редко, что, вполне возможно, связано с отсутствием ярко выраженных отличий от интервью в федеральных СМИ. На «втором» месте по количеству исследований находится интервью в немецком еженедельнике «Der Spiegel», а абсолютное «лидерство» в этой предметной области занимает интервью на YouTube и стратегии журналиста Юрия Дудя, самого известного интервьюера в русскоязычном сегменте. После того, как Юрий Дудь 3 года назад создал свой канал, молодые журналисты, вдохновившись его сюжетами, буквально заполонили видеохостинг YouTube подобными видеоинтервью с разными людьми: музыканты, политики, режиссёры, а также простые люди, в которых зрители узнавали себя. При этом каждый интервьюер старался первым пригласить интересную персону и подготовить всё более оригинальные вопросы. Таким образом, можно говорить, что интервью в течение последних нескольких лет возрождается.

Жанровая структура текстов СМИ впервые была сформулирована А.А. Тертычным. Интервью относится к информационным жанрам, т.к. его цель – узнать новую информацию у интервьюируемого о нём самом или о конкретном событии. «То обстоятельство, что на первый план в наши дни выходит информационная группа жанров, к которым относится и интервью, является первым веским аргументом в пользу его популярности в настоящее время» (Исакова 2009: 19).

В русскоязычных научных работах существует огромное количество определений понятия интервью, но определение, которое вводит С.Н. Ильченко, кажется нам наиболее полным, т.к. в нём интервью рассматривается с точки зрения взаимодействия участников общения. Итак, «интервью – акт коммуникации, предполагающий диалогическое общение журналиста с респондентом в ситуации последовательного чередования вопросов и ответов,

с целью получения информации, мнений и суждений, представляющих общественный интерес» (Ильченко 2003: 10).

Под словом «интервью» также понимается метод сбора информации и жанр, предполагающий оформленный в виде беседы аудио-, видео- и текстовый материал. В первом случае цель журналиста – собрать и классифицировать ответы респондентов, во втором – раскрыть личность известного человека и создать интересный контент для аудитории. Умение расположить гостя к беседе, оставаясь при этом «в тени», требует от интервьюера высокой квалификации и знания специфики жанра.

«Интервью – сложная срежиссированная система, отличающаяся специфическим характером постановки» (Багдасарян 2016: 16-17). В данном параграфе мы рассмотрим основные компоненты, составляющие эту систему.

Различные классификации интервью предлагали В.П. Пельта, А.А. Тертычный, А.А. Грабельников и другие, но все исследователи так или иначе сходятся в одном: интервью – это беседа. «На то, что интервью является беседой, указывает примерно одинаковый объем реплик журналиста и собеседника. В обычном же интервью на ответы собеседника отводится 70-90% текста» (Колесниченко 2008: 34).

То, что интервью не является повседневным диалогом, в котором равнозначное активное участие принимают оба коммуниканта, указывает:

1) степень гласности;
2) наличие инсценируемой ситуации;
3) цель коммуникации – раскрытие актуальной проблемы или создание объективного образа интервьюируемого;
4) координация хода беседы с помощью подбора гостей и выбора вопросов;
5) вопросно-ответный комплекс;
6) двуадресность – «информирование и воздействие не только на участников диалога, но и на читателя/зрителя» (Багдасарян 2016: 17; Иванова 2009: 9-10).

Интервью может протекать в двух формах: в виде монолога, где количество вопросов журналиста сведено к минимуму (иногда их вообще нет), и в виде беседы, когда журналист не только задаёт вопросы, но и сам выступает с рассуждениями (Колесниченко 2008).

Языковые особенности жанра интервью. И.В. Иванова, исследовав языковые единицы в различных интервью с 2002 по 2009 год (диктофонные записи, телеинтервью, тексты газетных интервью, интервью региональных и федеральных СМИ), в своей диссертации приводит список языковых средств, возникающих при трансформации интервью из устной формы в письменную (Иванова 2009). Перечислим некоторые из них:

- повтор как уточнение сказанного: *Я учился в Белорусской консерватории / **потом московскую аспирантуру закончил // Аспирантуру Московской консерватории** //;*
- наличие приёма эха в вопросно-ответном комплексе: *А.: **И этот театр…** Б.: **И этот театр…** / да // Он очень хотел поддержать его //;*
- большое количество разговорной лексики;
- минимальное присутствие деепричастных и причастных оборотов;
- инверсивные конструкции: *Б.: Бригада Пети была серьёзная / балет получил госпремию в России / и **Маску золотую** //;*
- использование незнаменательных или частично десемантизированных слов: Б.: *Ну как например / **скажем** / людей сгоняли в колхоз… //* (Иванова 2009: 10).

Расшифрованный текст интервью несколько раз редактируется: в изложении журналиста удаляются стилистические недочёты, повторы, слова-паразиты – речь участников приобретает чётко выстроенную композицию. «Печатная версия, как правило составляет около 1/3 объёма устного интервью» (Иванова 2009: 10).

Также исследовательницей было установлено, что в региональных письменных интервью из специальной лексики превалирует иноязычная и

сниженная: на 1000 слов приходится около 13 слов иноязычной лексики и 26 терминологической. А излюбленным средством выразительности у региональных и федеральных журналистов является эпитет, реже используется гипербола, метафора, метонимия и др. (Иванова 2009: 10).

Стандартность вопросов («*Расскажи немного о своём родном городе»*), молодёжный сленг (*«Хороший мотиватор»*) и фразеологические обороты («*набирать обороты»*) – те лексико-стилистические особенности, которые свойственны интервью (Лободенко 2016: 34).

О.О. Никифорова в ряду грамматических особенностей этого жанра подчёркивает номинативный характер изложения, который проявляется в большом количестве существительных и прилагательных; семантический повтор и рассуждение как доминирующий тип речи. «Для речевой формы “рассуждение” характерно также употребление придаточных уступительных с союзом “хотя”; придаточных условия с союзами “если... то”, а также придаточных причины с союзом “потому что” и изъяснительных придаточных» (Никифорова 2013: 64-73).

Жанр интервью в региональных изданиях имеет свою специфичность: он отвечает тематике издания и возрастным особенностям читательской аудитории, в том числе способствует сохранению языка как средства коммуникации в межэтническом многообразии представленного региона» (Бокова 2018: 172). Сохранение культуры, памяти и национального языка – важнейшая задача для народа Саха. Поэтому в интервью в прессе Якутии достаточно часто поднимаются такие проблемы, как влияние глобализации на молодёжь севера и воспитание патриотизма соответственно, популяризация языка, спортивных состязаний и местного ремесла. Региональный характер интервью проявляется в прежде всего в выборе темы.

«Стоит отметить, что внимание читательской аудитории к той или иной печатной прессе, в том числе к национальным изданиям, держится благодаря разнообразию публицистических жанров. Одним из ведущих жанров периодической печати, в том числе национальной, является жанр интервью»

(Щитова 2012: 34). Л.К. Лободенко, изучив медиатексты Челябинских СМИ за период 2010-2015 год, сделала вывод, что достаточно активно массмедиа в своей практике используют информационное сообщение (заметку), интервью и репортаж (Лободенко 2016: 33).

На лексическом уровне в региональном тексте наблюдается перевес в сторону заимствованной лексики, регионализмов, национально-маркированной лексики, редко диалектизмов. При этом, как показал анализ интервью в якутском электронном издании ЯСИА, вышеперечисленная лексика никак не маркируется. Предполагается, что читатель проживает или проживал когда-то в республике и ему понятны номинации, обозначающие уникальные предметы быта, явления природы и культурные достопримечательности.

Что касается грамматических особенностей, то они чаще всего проявляются в устной форме: в радио- и телеинтервью. Если интервьюируемый – представитель коренного народа Севера, а интервью проходит на русском языке, часто под влиянием родного языка возникают ошибки, заминки, междометия (среди них выделяют т.н. «тянущиеся гласные»), несогласованность частей речи.

Так как интервью, несмотря на широкий канал гласности, предполагает всё-таки доверительную обстановку: интервьюер должен расположить к себе гостя, а гость, в свою очередь, старается понравиться зрителю – у интервьюируемого есть возможность словом повлиять на аудиторию (выразить любовь к Родине, призвать к патриотизму и к сохранению самоидентичности). В первую очередь этой возможностью пользуются первые лица республики и известные патриоты, у которых берут интервью.

Доверительная беседа и стремление гостя показать себя искренним только способствуют речевому воздействию. О.В. Елецкая основными достоинствами интервью как предмета лингвистического исследования считает «возможность наблюдения за психологическими реакциями

опрашиваемого, личный контакт журналиста с респондентом, призванный обеспечить максимально полно реализацию вопросника» (Елецкая 2012: 75).

Теперь опишем основные виды интервью, т.к. использование конкретных языковых приёмов во многом будет зависеть от темы и цели интервью. В зависимости от темы выделяют – ***предметное, личностное и предметно-личностное интервью***.

Тема предметного интервью – положение вещей в какой-либо сфере. Например, предметное интервью часто берут у экономистов, когда нужно прокомментировать курсы валют или дать экономический прогноз на ближайшие месяцы. Частная жизнь эксперта не интересует журналиста. В региональной прессе предметное интервью встречается нечасто. В основном рекомендации, прогнозы и слова очевидцев публикуются в форме заметки, а не детального интервью.

Структура предметного интервью: «Вначале показывают связь собеседника с темой. Затем идет разбор предмета с целью его представления, прояснения и оценки. Завершается предметное интервью упорядочением сказанного, подведением итогов, практическими рекомендациями читателям» (Колесниченко 2008: 38).

Темой личностного интервью, как следует из названия, является человек: звёзды эстрады, спортсмены, художники и ремесленники. Задача журналиста – преодолеть «фасад» личности, «очеловечить гламурный образ» (Колесниченко 2008: 39). Беря личностное интервью, журналист иногда может совершать профессиональные ошибки: например, начать навязывать собственные оценки или политические взгляды, открыто демонстрировать пренебрежение к гостю, если он ему неприятен.

Это максимально свободный тип интервью, в котором нет чёткой структуры, а тема вопросов может быть ограничена героем интервью. Оптимальным считается подход, когда чередуются темы и типы вопросов (с открытым или закрытым ответом).

В ***предметно-личностном интервью*** рассматривается конкретный человек вместе с каким-то делом, которым он занимается. Это может быть учёный, совершивший открытие, писатель, опубликовавший один или несколько бестселлеров, олимпийский чемпион и др. Интервью такого плана обладают сильным воздействующим эффектом. Рассказывая о своих успехах, приглашённые гости мотивируют зрителей усердно работать, верить в себя. Это сближает предметно-личностное интервью с коучингом (англ. *coaching* – *тренировка*). Человек, достигший успеха, вместе с журналистом пытаются ответить на вопрос: *как этого достичь?* Собеседника последовательно расспрашивают о том, как он двигался к своему сегодняшнему состоянию, акцентируя внимание на напряженных и комических эпизодах, которые встретились на пути к успеху (Колесниченко 2008).

В зависимости от цели также выделяют:

- ***информационное интервью,*** рассчитанный на сбор материала для новостей (что? кто? где? когда? почему? зачем?)
- ***оперативное интервью*** – то же, что информационное, только в ещё более сжатом виде;
- ***опрос на улице или street talk*** – сбор мнений по какому-либо конкретному вопросу;
- ***интервью-расследование*** проводится с целью глубинного изучения какого-либо события или проблемы;
- ***интервью-портрет*** – то же, что и *личностное* в классификации по темам.
- ***креативное интервью*** воплощается в беседе, результатом которой является художественный очерк, документальный фильм, диалог в эфире (Лукина 2012: 17-18).

А.И. Кутний на примере материалов двух южных телеканалов предлагает классифицировать виды и типы интервью именно для региональных СМИ с учётом специфики их работы и структуры вещания.

Данную классификацию составляют: интервью-портрет, интервью для получения фактов и для выяснения мнения собеседника (Кутний 2012).

Таким образом, в интервью проявляется не только функция информативности – узнать больше о человеке или событии, но и функция воздействия: как в процессе беседы, так и на аудиторию. Исходя из цели и ситуации общения выбирается стратегия, используются приемы и тактики.

1.2. Основные классификации вопросов

Под интервью традиционно понимается диалог между интервьюером и интервьюируемым. Интервью имеет вопросно-ответную структуру и сочетает в себе несколько высказываний. Эти тематически взаимосвязанные высказывания внутри диалога (например, вопрос и ответ) называются диалогическим единством. «Вопросно-ответная серия – это диалог, включающий в себя более двух диалогических единств» (Багдасарян 2016: 16). Соответственно, диалогическое единство находится в отношении к диалогу как «часть-целое».

Диалогическое единство (термин Н.А. Купиной) – это структурно-семантическая единица диалогического текста, состоящая из двух и более компонентов (встречных высказываний участников диалога), «примыкающих к единому смысловому центру и взаимообусловленных семантически, структурно и коммуникативно» (Блох 1992: 10). «По количеству входящих в их состав компонентов диалогические единства подразделяются, соответственно, на простые (двухкомпонентные) и сложные (состоящие из трех, четырех и более взаимосвязанных компонентов). Своеобразие взаимообусловленности двух соседних реплик позволяет подразделить простые диалогические единства на три типа: повествовательно-отзывные, вопросно-ответные и побудительно-отзывные» (Есенина 2014: 135). Как уже было сказано ранее, для нашей работы ключевыми являются вопросно-ответные диалогические единства.

Компоненты диалогического единства у разных авторов называются по-разному: реплика-акция – реплика-реакция (Святогор), управляющая реплика

– зависимая реплика (Леонова, Шубин), стимул – реакция (Арутюнова), зачинный компонент – конечный компонент (Блох, Поляков). Мы будем пользоваться терминами *стимул – реакция* Арутюновой, т.к. в интервью вопрос является неким раздражителем, тем самым стимулом, который провоцирует реакцию (ответ) человека.

Согласно «Русской грамматике», ***вопросительными*** называются предложения (далее – ВП), «в которых специальными языковыми средствами выражается стремление говорящего узнать что-л. или удостовериться в чем-л. Вопросительные предложения, таким образом, информируют о том, что хочет узнать говорящий» (Русская грамматика 1980). Главная цель ВП – запрос информации.

Узнать вопросительное среди других предложений помогает знак вопроса, вопросительное местоименное слово *(кто, что какой, чей)*, вопросительные частицы *(ли, не… ли, неужели, что если)* и в устной речи – интонация. «С помощью этих средств любое невопросительное предложение может стать вопросом или переспросом» (Русская грамматика 1980).

«Вопросительные предложения часто имеют субъективно-модальную окраску» (Русская грамматика 1980). В них может выражаться отношение говорящего: неуверенность, предположение, довольство или недовольство и другие значения. Выбор языковых средств в вопросах интервью будет зависеть от социальных и психологических особенностей коммуникантов: возраст, образование, пол, темперамент – и от ***коммуникативной ситуации***. Коммуникативная ситуация включает участников общения и все условия, в которых оно протекает. От коммуникативной ситуации во многом зависит формулировка вопросов. Компонентами ситуации общения являются: адресант (кто?) и адресат (кому?), цель общения (зачем?), сообщение и тематическое содержание (о чём? какова тема?), язык как средство общения, сфера и условия речевого акта (научная, учебная, бытовая, профессиональная, официально-деловая, социально-культурная), тактики и стратегии речевого поведения. Дополнительными характеристиками может быть

культурологическая составляющая, этикетные нормы и внеязыковые факторы (например, присутствие других участников общения и обстановка беседы).

Чтобы интервью прошло удачно, журналист должен хорошо подготовиться: узнать побольше информации о госте, написать опорные вопросы исходя из прагматических целей, нужным образом настроиться. Тем не менее, строго придерживаться плана интервью не всегда возможно: большая часть вопросов импровизируются. «Именно формулировка вопросов является ключевым этапом при подготовке интервью. Интервьюеру следует избегать чрезмерной академичности и специализации вопросов, они должны быть понятны как респонденту, так и слушателю (читателю). Кроме того, респондент имеет свои сильные стороны, к которым он будет пытаться привлечь внимание, и недостатки, которые он, естественно, попытается скрыть» (Коготкова 2013: 102). Не следует также одновременно задавать два и более вопроса. В этом случае респондент ответит только на один, причём на самый простой (Коготкова 2013: 102)

Интервью может быть стандартизованным с чётко поставленными вопросами и неформализованным. Вопросы – не сама цель интервью. Они направляют мысль гостя в нужную интервьюеру сторону, помогают раскрыть его личность и получить новую информацию. Бестактными, слабыми, уже наскучившими, слишком напористыми или слишком общими вопросами можно «отпугнуть» героя и провалить интервью.

В журналистской практике все вопросы делят на ***открытые и закрытые***. Закрытый вопрос предполагает ответ либо «да», либо «нет». Закрытый вопрос включает в себя подвиды: альтернативный, с отрицанием, прямой или косвенный. Открытый вопрос требует развёрнутого ответа, не ограниченного по объему. Основные достоинства открытого вопроса заключаются в следующем:

- он стимулирует к диалогу, даёт возможность гостю быть повествователем, свободно рассказывать о своих чувствах, комментировать события;

- ориентирует человека на размышления, анализ своих поступков, стимулирует рождение мыслей, которые ранее, может быть, и не приходили ему в голову.

Недостатками могут быть:

- пространный ответ, «поток сознания», сложный для понимания;
- необходимость вмешательства журналиста с целью возвращения к главной теме, что может обидеть героя;
- смущение гостя, который не привык отвечать на общие вопросы или не доверяет интервьюеру;
- высокая затрата времени (Лукина 2012 : 20).

Исходя из типа вопроса выделяют диалогические единства с открытыми вопросами и диалогические единства с закрытыми вопросами (Есенина 2014: 136-137).

Чтобы правильно понять, что имеет в виду интервьюируемый, журналист может задать ***уточняющий вопрос***. Ответив на него, гость привносит больше информации, что может в дальнейшем избежать двусмысленности. Также многие вопросы в интервью по форме являются ***стимулирующими комментариями,*** а не вопросительными предложениями (Коготкова 2013: 101). Как правило, люди не могут говорить на все темы свободно. Возможно, от смущения, страха показаться глупыми, стремления что-то скрыть, они замолкают, жмутся, начинают заикаться. Тогда интервьюер может сказать: *«Пожалуйста, продолжайте, это очень интересно…»* или в коротко повторить ранее сказанное и задать ***развивающий вопрос***. В научной литературе встречаются разные названия для таких реплик: стимулирующий комментарий, развивающий вопрос, реплика-стимул.

Другой тип вопроса – ***контрольный***. Он задаётся в том случае, если у журналиста есть сомнения в достоверности слов интервьюируемого. «Будучи разновидностью контрольного, вопрос уличающего характера применяется в случае явных противоречий в ответах, а также если собеседник был

непоследовательным в описании, неуверенным в аргументации» (Лукина 2012: 23).

Выделяют и более частные типы вопросов:

- количественные – вопросы с количественным местоимением, призванные дать числовую характеристику;
- гипотетические – вопросы, побуждающие собеседника порассуждать о событиях в будущем или прошлом, дать прогноз, пофантазировать о перспективах или последствиях;
- переходные – «вопросы-переключатели», чтобы изменить тему беседы, когда она исчерпана.

У всех типов вопросов есть своя специфика, когда и кому их следует задавать. Например, развивающие вопросы лучше задавать после закрытых, на гипотетические, скорее всего, не будут отвечать политики и люди, часто принимающие решения; переходные вопросы должны быть достаточно интересными, чтобы гость захотел «переключиться» на новую тему.

Особое внимание при подготовке к интервью отводится первому вопросу. Первый вопрос может задать тон всего интервью. Как правило, он связан с обыденными вещами (отдых, развлечения, внешний вид), личной жизнью (переживаниях, волнующих проблемах), профессией и родом деятельности (Шатохина 2006: 62). Его цель – расположить к себе интервьюируемого и заинтересовать аудиторию, которая потом будет это интервью смотреть или читать.

Теперь обратимся к собственно лингвистической классификации ВП, описанной в «Русской грамматике» (§ 2623), с точки зрения функциональной семантики. ВП объединяются на основе первичных и вторичных функций вопросов. К первичной относится главная функция вопроса – запрос информации; ко вторичной – передача позитивной информации, всегда эмоционально окрашенной.

1. Первичные функции устанавливаются в зависимости:

1.1. От характера и объема той информации, которая должна быть получена:

1.1.1. ***Общевопросительные*** – направлены на получение информации о ситуации в целом *(Что случилось? Что ему нужно? Что вы думаете об этом человеке?);*

1.1.2. ***Частновопросительные*** **–** заключают вопрос об отдельной стороне какого-л. факта, о деятеле, носителе состояния, тех или иных обстоятельствах *(Это вы делали? Как хорошо вы его знаете?)*.

1.2. От осведомлённости говорящего о том, что спрашивается:

1.2.1. ***Собственно-вопросительные*** отражают полную неосведомлённость спрашивающего *(Другая дорога тут есть? Кто это сделал?);*

1.2.2. ***Неопределённо-вопросительные*** совмещают вопрос с догадкой, предположением, неуверенностью, сомнением *(Вы не меня ждёте? Вы как будто расстроены?);*

1.2.3. ***Констатирующе-вопросительные*** – вопрос с почти полной уверенностью *(Значит, я не ошибся? Ведь вы – офицер, да? Ведь у тебя билет уже есть?).*

1.3. От ожидаемого ответа:

1.3.1. ***Предложения, требующие ответа-подтверждения или ответа-отрицания*****,** т.е. ответа о соответствии или несоответствии действительности *(Тебе больно? Существуют акты или не существуют? Он спит?);*

1.3.2. ***Предложения, требующие в ответе информации, сообщения о том, что спрашивается*** *(А почему ты плачешь? – Потому что радуюсь, что такие люди, как твой папа, бывают на свете.).*

2. Вторичные функции выражаются в следующих случаях:

2.1. Вопрос, в котором заключено уверенное экспрессивно окрашенное утверждение *(Это не самоуправство ли? Разве это не красота? Да ты должна, что ли?);*

2.2. Вопрос, в котором заключено уверенное экспрессивно окрашенное отрицание (риторический вопрос – *Кто себе зла желает?*);

2.3. Вопрос уяснение, повторяющий словесный состав предшествующей реплики и обычно осложненный эмоциональной окраской удивления, недоумения, беспокойства, неодобрения:

2.3.1. ***Предшествующая реплика – вопросительная*** *(Отчего же ты такой, словно с цепи сорвался? – Я с чего такой?; Чего ты хочешь? – Чего я хочу?);*

2.3.2. ***Предшествующая реплика – невопросительная*** *(Да что же делать? – Бежать. – Бежать?).*

2.4. Вопрос – побуждение к чему-л.:

2.4.1. ***Вопрос, побуждающий к действию*** *(А впрочем, чего же мы стоим?);*

2.4.2. ***Вопрос, побуждающий к прекращению действия*** *(Да бросите ли вы в конце концов вашу музыку?);*

2.5. Вопрос, выражающий эмоциональную реакцию говорящего: констатация факта, оценка, отношение, аффективное состояние *(Да что вы смеетесь надо мной? Как вам не совестно?);*

2.6. Вопрос, имеющий целью активизировать внимание, заинтересовать, обратить внимание на форму выражения мысли *(И что ж бы вы думали? Как бы это проще объяснить?)* (Русская грамматика).

Функциональный аспект в интервью проявляется наиболее ярко: интервьюер беседует с респондентом, оказывает на него речевое воздействие, для достижения коммуникативной цели (Коготкова 2013: 101). Исходя из этой цели интервьюером подготавливаются вопросы. Данная классификация ВП с точки зрения их функций отражает интенцию говорящего и является достаточно полной, т.к. учитывает объём запрашиваемой информации, предшествующие реплики и ожидаемые ответы.

В тексте интервью главным структурным элементом является заголовок. «Заголовки-вопросы являются характерной чертой современных публицистических текстов интервью <…> Прагматический эффект вопроса в заголовке статьи заключается в том, что он подсознательно подталкивает читателя к поиску ответа на него, но для того, чтобы ответить на вопрос, нужно прочитать и осмыслить следующий за ним текст» (Бычковская 2016: 60). Ставя в заголовок вопрос, автор как бы вступает в диалог с читателем, обращает внимание на проблему и предлагает также над ней задуматься. Наиболее частой структурой вопросительных заголовков в публицистических интервью являются высказывания с вопросительными словами и вопросы, содержащие обращения (Бычковская 2016: 60).

Как мы убедились, одна из трудностей в обработке естественного языка состоит в вариативности классификаций вопросов: их группируют согласно логической структуре, предполагаемым ответам, целям, речевым стратегиям и тактикам, которые они реализуют. Многообразие ВП представляет собой проблему при создании ***вопросно-ответных систем*** *(от англ. QA – Question-answering system)* – информационной системы, способной принимать вопросы и отвечать на них (далее ВОС).

Для моделирования ВОС первоочередной является задача определения вопроса с учётом его функции в диалоге и семантики, чтобы в дальнейшем система из базы знаний (электронного хранилища) могла найти нужную информацию по ключевым словам. На этапе анализа запроса возникает проблема, которая состоит в том, что по формальному строению ВП или совпадают с невопросительными, или строятся по собственным синтаксическим образцам (Русская грамматика 1980). Одну и ту же информацию можно запросить разными способами, в связи с чем важно корректное распознавание семантики слов, независимо от идиом, омонимии и используемой лексики. Чем лучше QA-система будет распознавать поступающий вопрос, тем эффективнее будет работать поиск и фильтрация документов и тем точнее будет вывод, конечный ответ.

1.3. Метод моделирования в прикладной лингвистике

Моделирование используют во многих науках. Под этим термином понимают метод, при котором реальный объект действительности или явление представляется в виде модели – схемы, уменьшенной копии или другой имитации. Он нацелен на выяснение способов функционирования реального объекта или его свойств. Модель используют в прагматических целях, чтобы посмотреть и предсказать, как поведёт себя реальный объект, сделать выводы, предупредить ошибки и избежать лишних расходов при дальнейшем реализации модели.

Метод моделирования активно применяют в лингвистике: например, при обучении иностранным языкам моделируется коммуникативная ситуация, а в прикладной лингвистике моделируются реальные лингвистические объекты, их системы, а также процессы речемыслительной деятельности человека. Впервые понятие лингвистической модели появилось в структурной лингвистике. В 60-70-х годах XX века термин стал активно использоваться в математической лингвистике вместе с другими математическими определениями. В последние годы возросло количество исследований, обозначивших своим основным методом моделирование или имеющих целью построение модели своего предмета, благодаря чему «география филологических проблем заметно детализировалась» (Белоусов 2010: 94).

Моделирование – это «исследование каких-либо явлений, процессов или систем объектов путём построения и изучения их моделей для определения или уточнения характеристик и рационализации способов построения вновь конструируемых объектов» (БЭС 1997).

Моделирование в лингвистике – это «составление схемы или модели какой-либо языковой единицы. *Моделирование сложного слова. Моделирование сложноподчинённого предложения с придаточным определительным»* (Розенталь 1976).

Широкое применение метода моделирования в различных сферах человеческой деятельности придало ему статус общенаучного метода познания.

Как пишет Ю.Д. Апресян, «моделирование возникает во всех тех научных областях, где объект науки недоступен непосредственному наблюдению» (Апресян 1966: 78). Учёный сравнивает построение модели с «чёрным ящиком» – «тот скрытый от исследователя механизм, который осуществляет переработку исходных материалов в конечные продукты» (Апресян 1966: 78). Модель – это посредник между теорией и практикой. В связи с тем, что модель выполняет гносеологическую и информационную функции, модель не является зеркальным копированием изучаемого объекта или явления, она несёт новое знание. Иными словами, механизм работы нам не известен, мы знаем только, какие материалы он получает «на входе» и какие конечные продукты он производит.

В сфере прикладной филологии сложился объёмный комплекс проблем и направлений (лингводидактика, политическая, юридическая и компьютерная лингвистика: автоматический перевод, распознавание речи, создание поисковых систем), требующих расширения границ лингвистической науки в сторону поиска новых междисциплинарных связей. «Методологическое значение моделирования для лингвистических наук определяется тем, что именно на принципах моделирования базируется практическая реализация методологических подходов к анализу основного объекта филологических наук – текста – и всех методов его исследования, основанных на использовании математического аппарата и средств вычислительной техники» (Морозов 2001: 36).

Таким образом, **модель** в лингвистике – это «искусственно созданное лингвистом реальное или мысленное устройство, воспроизводящее, имитирующее своим поведением (обычно в упрощенном виде) поведение какого-либо другого («настоящего») устройства (оригинала) в лингвистических целях» (ЛЭС 2002). Модель уподобляется реальному

объекту, устройство которого нам хорошо известно. В процессе познания действительности модель выполняет ряд функций: замещение моделируемой системы; информационную; гносеологическую; формализационно-алгоритмическую; доказательственно-иллюстративную.

Ю.Д. Апресян выделяет три вида моделей, отличающиеся друг от друга характером рассматриваемого объекта:

- модели речевой деятельности человека, имитирующие конкретные языковые процессы и явления;
- модели лингвистического исследования, имитирующие те исследовательские процедуры, которые ведут лингвиста к обнаружению того или иного языкового явления;
- метамодели, имитирующие теоретическую и экспериментальную оценку готовых моделей речевой деятельности или лингвистического исследования (Апресян 1966).

Более частная классификация лингвистических моделей представлена на рисунке 1:

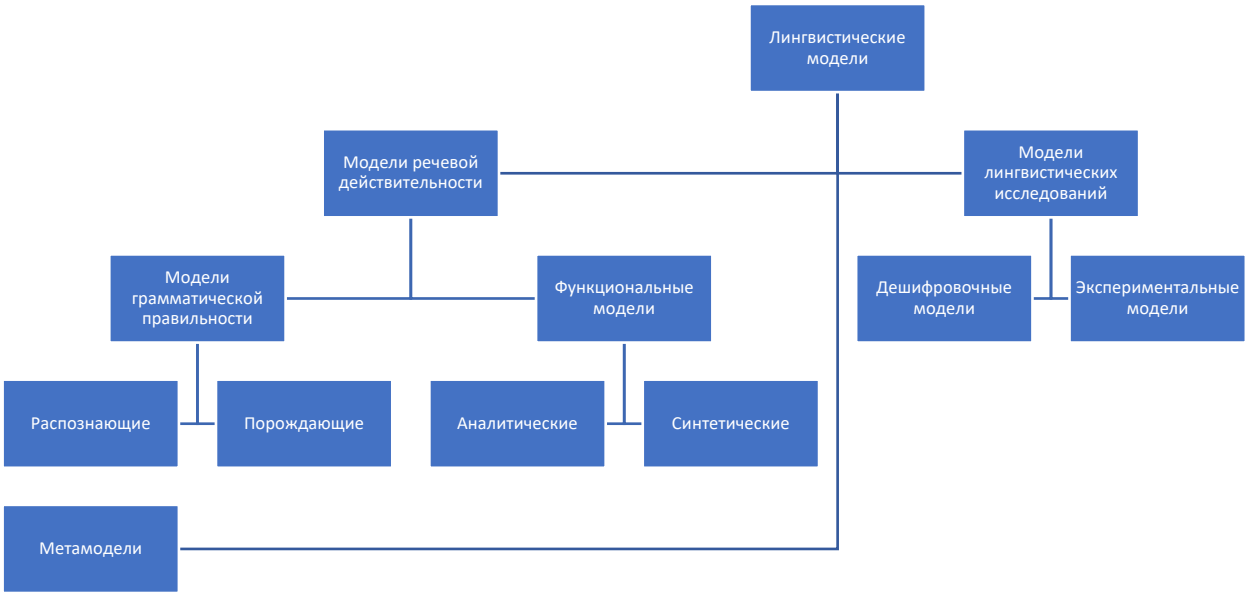

Рисунок 1 – Лингвистические модели (Морозов 2001: 38)

Так как речевая деятельность человека является сложнейшим актом, модели, её отражающие представляют собой сложную и разнообразную структуру. Рассмотрим виды лингвистических моделей отдельно и кратко обозначим предмет, который заменяет та или иная модель.

1. **Модель грамматической правильности** имитирует умение отличать правильное от неправильного в языке. По типу информации они подразделяются на распознающие и порождающие.

Распознающие модели на «вход» получают некоторый отрезок текста на естественном языке и выдают информацию, является ли данный отрезок грамматически правильным. *Пример: программы, проверяющие корректность написанного кода.* Порождающие, наоборот, на «входе» имеют совокупность правил и набор минимальных единиц, а на «выходе» – сконструированное высказывание. *Пример: грамматика Хомского.*

2. **Функциональные модели** имитируют умение соотносить план содержания с планом выражения. По тому, какой аспект речевой деятельности моделируется, функциональные модели делят на аналитические и синтетические. Аналитическая модель на «вход» получает некоторый отрезок текста и даёт на «выходе» его семантическое представление на специальном семантическом языке. *Пример: учение о структурных типах простого предложения (N_1 – Vf).*

Синтетические модели обратны аналитическим: на «входе» принимается семантическая модель, по которой строится множество синонимичных текстов, выражающих этот смысл (Морозов 2001: 37-39).

«Модели анализа и синтеза – необходимая часть модели перевода, в том числе и автоматического, и различных систем искусственного интеллекта» (Морозов 2001: 39).

Метамодели и модели исследования второстепенны по отношению к моделям речевой действительности, т.к. они выполняют вспомогательную роль.«Метамодели, разрабатываемые математической лингвистикой, представляют собой математические теории, объектами которых являются не

отдельные лингвистические понятия, а целостные модели языка» (Морозов 2001: 40).

3. Дешифровочные модели исследования используют ограниченный корпус текстов, все сведения о языке модель должна извлечь только из текстовых данных.

4. В экспериментальных моделях задаётся множество правильных текстов данного языка; по ходу эксперимента лингвист прибегает к помощи информанта (носителя языка).

Требования к модели. Особенность моделирования состоит в том, чтобы создавать модель, максимально приближённую к реальности. Для этого в исследовательской деятельности должна быть (1) цель, (2) определённый набор методов, приемов и принципов и (3) некоторый идеал.

Цель конкретизирует *полноту* модели – количество фактов, необходимых для точного представления модели. Включение в модель избыточных средств, не отвечающих цели исследования, перегружает модель и делает её неудобной. Поэтому в моделировании следует использовать то количество средств, которое необходимо для достижения научной цели. Этим достигается *простота* использования и *экономичность* в расходовании энергетических и временных ресурсов при применении модели.

Наличие идеала определяет *точность* – возможность выполнения операций с помощью модели и *адекватность* – максимальную похожесть с моделируемым объектом.

Но главное требование к модели заключается в том, что она должна не только схематично представлять какой-либо онтологический языковой объект, но и, как подчёркивают многие исследователи, генерализировать новые знания об объекте. «Под моделью понимается такая мысленно представляемая или материально реализованная система, которая, отображая или воспроизводя объект исследования, способна замещать его так, что ее изучение дает нам новую информацию об этом объекте» (Штофф 1966: 19).

В модельной лингвистике актуален системнодеятельностный подход, который состоит в «системе положений, позволяющих описывать, нормировать саму исследовательскую деятельность лингвиста. Системнодеятельностный подход как основа модельной лингвистики может «усиливаться» рядом других подходов и установок, в частности, экспериментальным подходом, квантитатичным и практикоориентированным характером создаваемых моделей. Структура модельной лингвистики кроме теоретико-методической основы должна включать методы вероятно-статистической обработки данных, экспериментальную лингвистику и компьютерную лингвистику» (Белоусов 2010: 96). При этом компьютерное моделирование следует понимать шире, чем «способ», облегчающий работу лингвиста. Применение компьютерного моделирования позволяет исследователям оперировать большими объёмами данных (текстовыми массивами) и обрабатывать их методами вероятностно-статистического анализа.

В лингвистике используется синтаксическое моделирование текста, которое формируется в рамках функционально-коммуникативного синтаксиса и теории текста; моделирование функции языкового значения, основанного на семиотике и когнитивной лингвистике, и моделирование лексической семантики. Методы семантического моделирования применяются практически в каждой информационной системе. В зависимости от задач в создаваемых приложениях компьютерной лингвистики семантика языка рассматривается в определённых ракурсах. Существуют подходы, направленные на моделирование структурной (синтаксической) составляющей языка вкупе со смысловыми единицами, и лексико-семантические методы, фокус которых смещён в сторону отдельных лексических единиц и взаимодействия их смыслов: тезаурусы, онтологии, семантические словари и т.д. (Кожунова 2012: 86). «При этом два принципиально разных полхода к моделированию языковых реалий и

процессов развиваются параллельно и дополняют друг друга» (Кожунова 2012: 86).

Лексико-семантический подход применяется для автоматического поиска терминов в интернете в качестве ответов на сложные вопросительные предложения. Данный метод основан на присвоении весов терминам, исходя из степени значимости в определённой предметной области. «Веса терминов в основном используются для отображения семантических связей между терминами и узловыми понятиями» (Кожунова 2012: 87). Это необходимо, чтобы предоставить пользователю наиболее релевантную информацию, т.к. делая запрос в поисковой системе, пользователю наряду с правильными результатами (отвечающими запросу) выдаётся массив ненужной, попутно найденной информации. Возможное решение данной проблемы видится в вопросно-ответном методе поиска, дополненном базой знаний с семантическими характеристиками терминов и весами для построения ассоциаций между ними.

1.4. Структура работы вопросно-ответной системы и примеры её реализации

В последние годы активизировались исследования в области человеко-машинного диалога, а именно, использование диалога как способа общения и как вида текста. Диалог между человеком и машиной означает обмен посланиями в установленной языковой форме для достижения коммуникативной цели. **Вопросно-ответная система** (от англ. Question Answering Systems) – это вид информационно-поисковых систем, способных обрабатывать введённый пользователем вопрос на естественном языке и выдавать осмысленный ответ.

«Прагматически определяющим фактором человеко-машинного диалога является то, что он формируется на основе связи темы диалога, ситуации и контекстов взаимодействия в зависимости от коммуникативных задач, стоящих перед участниками общения, а также языковых возможностей, которыми обладают участники диалога» (Сулейманов 2016: 24). Выявление и

учёт этих прагматических характеристик влияет на распознавание вопроса и, как следствие, точность выдачи ответа. «В общем случае задача построения вопросно-ответной системы включает в себя создание механизма сбора данных, построение собственной базы данных, и разработку интерфейса выдачи результата» (Кошкаров 2020: 203). База данных представляет собой собранные из различных источников материалы.

Диалоговое взаимодействие пользователя с машиной протекает в одном из трёх режимов: 1) когда на вопросы системы отвечает пользователь, 2) когда на запрос пользователя определённым образом реагирует система, и, наконец, 3) двухсторонне активный диалог, когда пользователь и система меняются ролями в ходе общения (Сулейманов 2016: 27). Наиболее изученным является второй режим, когда пользователь задаёт вопросы и получает ответ или из базы данных, или он генерируется самостоятельно. Режимы 1 и 3 пока мало изучены.

Большинство лингвистических процессоров для общения с базой данных функционируют в предположении, что человек инициирует диалог, а не система. К таким работам относятся экспериментальная система TIBAQ (Text-and-Influence Based Answering of Questions), система SAM, ПОЭТ и ИВОС, Лингвистический процессор для сложных информационных систем (Ю.Д. Апресян), система InterBASE.

Существует множество различных подходов к организации архитектуры ВОС. Архитектуру типичной QA-системы можно представить в виде следующей схемы (см. рисунок 2):

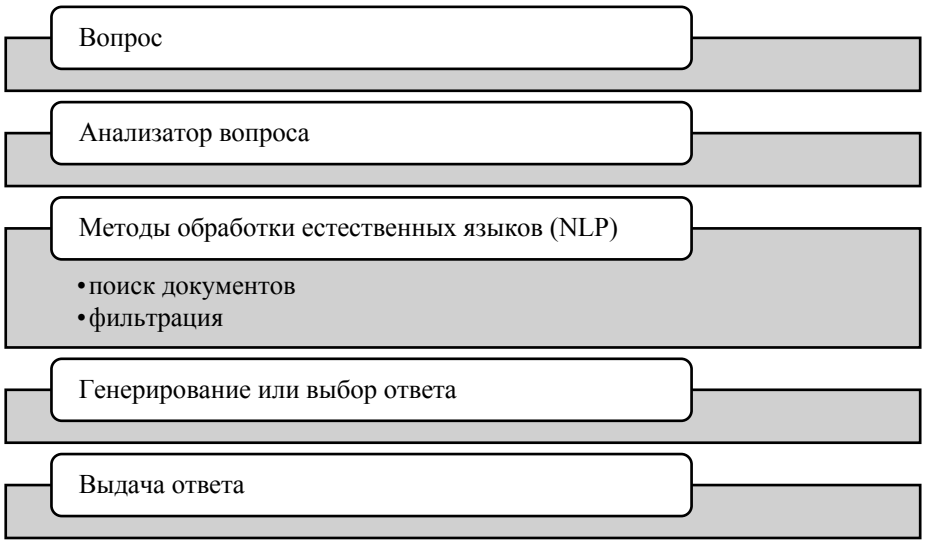

Рисунок 2 – Архитектура QA-системы

Как показано на схеме, ВОС состоит из пяти блоков. Рассмотрим подробнее их алгоритмы работы.

1) Вопрос. На вход система получает от пользователя вопрос-запрос, который задаётся на естественном языке. Структура вопроса накладывает ограничения на форму ответа и его содержание.

2) Анализатор вопроса. Для понимания семантики вопроса существуют различные подходы: выделение шаблона вопроса с помощью регулярных выражений; семантическая типизация вопросов и ответов, когда каждому типу вопроса противопоставляется смысловая конструкция (формула); выделение фокуса и опоры вопроса; присвоение семантических тэгов заранее заданной таксономии; использование морфологических шаблонов и синтаксических шаблонов (вопросительное предложение представляется в виде синтаксического дерева). Далее мы рассмотрим их поподробнее.

3) Методы обработки естественного языка. Во многих системах анализ ответа в тексте сводится к тому фрагменту, где содержится ответ на заданный вопрос. Выделенная часть текста подвергается лингвистической обработке: исправляются различные ошибки, удаляются ненужные символы и прочее; текст разбивается на предложения; предложения, в свою очередь, на слова; производится разбор всех слов (морфологический, синтаксический, семантический).

4) Генерирование или выбор ответа. Далее, используя базу результатов автоматической обработки создается запрос, передаваемый поисковой машине, которая выбирает документы, наиболее удовлетворяющие запросу. Источником информации является локальное хранилище с документами (база знаний) или интернет. В первом случае возникает необходимость хранения локальной копии информации, во втором – проблема доступа к внешним ресурсам.

Если в качестве источника используется база знаний, текст каждого подходящего документа подвергается обработке, выделятся список возможных кандидатов на ответ. Если источником служит интернет, то извлекаются потенциально релевантные веб-ресурсы с самым высоким рейтингом.

Если задачей является автоматический поиск терминов, тогда из полученного корпуса документов удаляется вся лишняя информация типа html-тегов, происходит обработка морфологическим анализатором для дальнейшего извлечения существительных. «Для извлечения простых и составных существительных вычисляется частота встречаемости в документах и в интернете. В результате словам присваиваются соответствующие веса. Простые и составные существительные ранжируются в соответствии с их весом» (Кожунова 2012: 90).

5) Выдача ответа. Определённые фрагменты текстов с ссылками на веб-сайты или без них, или список терминов предоставляются на выход (в качестве ответа). От ответа, в свою очередь, требуется раскрытие определённого смысла и вывод необходимого объёма информации, запрашиваемого в вопросе.

Остановимся на втором этапе – анализ вопроса – и рассмотрим наиболее известные методы анализа семантики и структуры вопроса.

«Извечная» проблема анализа вопроса состоит в выделении его главного элемента – семантики, сути. Многообразие форм представления смысла

требует выделения типовых смысловых конструкций, основанных на «жесткой грамматике», шаблонах или статистических подсчётах.

Для распознавания семантики вопросов на естественном языке может использоваться метод, при котором выделяется *фокус* вопроса, *опора* вопроса и *семантический тэг* ответа.

Фокус вопроса – это вопросное словосочетание, «такие сведения, содержащиеся в вопросе, которые несут в себе информацию об ожиданиях пользователя от информации в ответе» (Соловьёв 2010: 43).

Опора вопроса – «это остальная часть вопроса (после «вычета» фокуса), которая несёт в себе информацию, поддерживающую выбор конкретного ответа» (Соловьёв 2010: 43).

Семантический тэг ответа – «класс запрашиваемой пользователем информации согласно некоторой ранее заданной таксономии» (Соловьёв 2010: 43). Например, в предложении «*когда начнётся распродажа в меге?»* семантическим тегом является Date, «*когда начнётся распродажа»* – это фокус вопроса, а фраза «*в меге*» – опора вопроса (см. таблицу 1). «Таксономия семантических тэгов обычно выбирается разработчиками системы так, чтобы покрыть большую часть вопросов к системе» (Соловьёв 2010: 44).

Таблица 1 – Примеры анализа вопросов из заданий РОМИП 2009 (Соловьёв 2010: 44)

№	Вопрос, жирным шрифтом выделен фокус	Семантический тэг
nqa2009_6368	**как отключить** перехват клавиатуры?	Recipe
nqa2009_7185	**сколько стоит** починить гнездо у телефона сони эрикссон?	Money
nqa2009_6425	в каких религиях **как рассматривается** карма?	Definition
nqa2009_3123	отечественная война **кто с кем** ?	Country
nqa2009_8557	**являются ли** чердаки пожароопасными помещениями?	Yes/No
nqa2009_7801	**какое колличество циклов** чтения/записи предусмотренно компанией fujifilm для картриджей стандарта lto 4?	Cardinal
nqa2009_8763	**когда начнется распродажа** в меге ?	Date
nqa2009_9150	**во сколько заход солнца** 27 февраля?	Time
nqa2009_8754	**когда можно сводить** кошек?	Age
nqa2009_6797	какие в тамбове есть **студии звукозаписи**??	Organization

При использовании семантических тегов, возникает следующая проблема: связь между вопросительными словами и семантическими тэгами не так прямолинейна. «Так, слово «кто» может свидетельствовать и о персоне,

и об организации, и о стране, и о народе (например, в вопросе *«Кто выиграл войну?»*).

Фокус и опора вопроса, а также семантический тэг выделяются в метапоисковых системах. В качестве источника данных такая система использует классическую поисковую систему, т.е. использует неструктурированные данные, и формулирует запрос по ключевым словам, входящим в опору вопроса (Воробьёв 2020: 146-147). «Достоинства такой системы заключаются в отсутствии необходимости хранить огромный массив информации (для поиска в интернете) и гибкости — система может использовать любые доступные инструменты для анализа фрагментов (поиск по ключевым словам, контекстный поиск, полнотекстовый поиск), представлять фрагменты в виде графов. Недостатки — высокая вычислительная нагрузка в момент обработки вопроса» (Воробьёв 2020: 148).

Для выделения фокуса вопроса используются шаблоны, включающие морфологическую информацию (см. таблицу 2).

Таблица 2 – Примеры шаблонов для выделения фокуса в английском языке (Abraham 2006)

Вопросительное слово	**Шаблон**
What, which , name , list, identify	question word + headword of first noun cluster
Who, why, whom, when	question word
Where	question word + main verb
How	question word plus next word if it seeks an count attribute + headword of first noun cluster
	question word plus the next word if it seeks an attribute
	if question seeks a methodology, then just question word

Простейшим методом извлечения целых вопросов из текста или фокуса вопроса является подготовка символьных шаблонов – **регулярных выражений**. Очевидный недостаток в использовании регулярных выражений заключается в том, что в речи вопросительные предложения отличаются от идеальных вопросных моделей, представленных в шаблонах. Как мы уже выяснили, порой вопрос задаётся в форме стимулирующего к ответу комментария, не обладает вопросительным словом или не выделяется

вопросительным знаком на письме. «Обойти» правила системы в таком случае очень легко.

Следующим шагом после символьных шаблонов идёт метод построения синтаксических деревьев – **синтаксический шаблон**. Исходной точкой данного метода является вопросительное слово и его связующие. В связи с тем, что фокус вопроса находится в определённом синтаксическом отношении с вопросительным словом, предлагается построить синтаксическое дерево предложения. Чтобы система успешно обработала каждое слово, предварительно она должна исправить ошибки и опечатки. ВП представляется в виде синтаксического дерева, к которому задаётся аннотация с морфологическими метками в узлах.

Синтаксический шаблон лежит в основе поиска по аннотированному тексту. Такие системы имеют в своём составе поисковый индекс документов и осуществляют поиск по неструктурированным данным. Элементами индекса являются не отдельные слова текста, а именованные сущности, элементарные синтаксические связки (пары грамматически связанных слов), предикативно-аргументные структуры предложения. «Построение индекса происходит с привлечением компьютерной лингвистики, а именно: каждый новый документ проходит автоматическую обработку на естественном языке, размечаются объекты вопросно-ответной системы, затем они добавляются в индекс» (Воробьёв 2020: 148).

Теперь рассмотрим современные QA-системы. Большинство существующих в настоящее время вопросно-ответных систем ориентированы на английский язык, тем не менее, русскоязычные вопросно-ответные системы сегодня имеют все шансы соперничать с аналогичными английскими системами.

Все вопросно-ответные системы делятся на две группы – узкоспециализированные и общие.

Узкоспециализированные (closed-domain) применяются в конкретных предметных областях: медицина, юриспруденция, справочная информация.

Общие (open-domain) осуществляют поиск ответов из любых областях знаний. Наиболее известная общая ВОС – это START. Это первая QA-система. Она была запущена в 1993 году для английского языка. «Система отвечает на вопросы, предварительно распределяя их по категориям: наука и справочная информация; искусство; география; история и культура» (Кошкаров 2020: 203).

START использует два подхода при поиске ответов: аннотация знаний и интеллектуальное извлечение знаний (START). Аннотация знаний используется, если есть проверенный источник, где можно найти ответ на вопрос-запрос. Если такую информацию найти нельзя, используется «интеллектуальное» извлечение знаний. В этом случае результат разбора вопроса переадресовывается поисковой системе. Полученные кандидаты комбинируются с учётом веса каждого. В результате генерируется окончательный ответ.

Одной из первых ВОС в начале 60-х годов была *BASEBALL*, отличительной особенностью которой являлась возможность задавать вопросы к системе на естественном языке. База знаний как обычно служила структурированная база данных. В этом заключался главный недостаток всех ранних ВОС: отсутствие *BigData* – большого объема оцифрованных фактов и правил.

Lasso – наиболее развитая ВОС, разработанная в лаборатории компьютерной лингвистики Южного Методического университета в США. Архитектура Lasso состоит из трёх основных модулей (см. рисунок 3):

1. модуль обработки вопроса:
2. модуль обработки абзацев;
3. модуль обработки ответа.

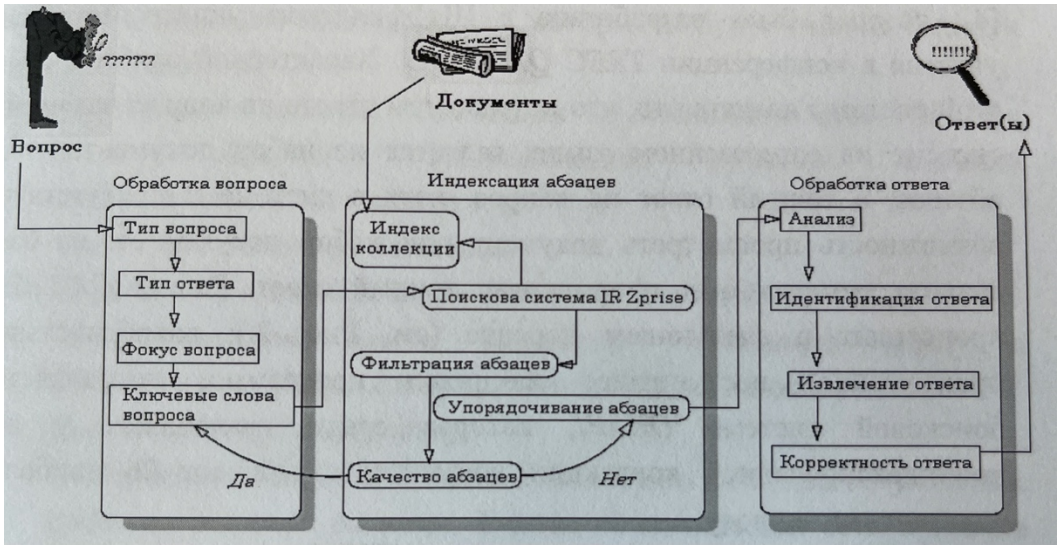

Рисунок 3 – Архитектура BOC Lasso (Захаров 2015: 5)

Модуль обработки вопроса определяет тип задаваемого вопроса («what-who», «what-when», «how-long» и т.д.), семантический тэг (DATE, LOCATION, PERSON) и фокус вопроса.

Модуль обработки вопроса также определяет ключевые слова запроса, которые должны быть переданы модулю индексации данных.

Модуль обработки абзацев производит оценку качества найденных абзацев. «В случае признания качества удовлетворительным, производится их упорядочивание в соответствии со степенью правдоподобия содержания ответа, в противном случае происходит добавление или удаление некоторых ключевых слов поиска, после чего поиск по обновлённому списку ключевых слов возобновляется и система повторно оценивает качество найденных абзацев. Модуль проверки качества абзацев позволяет разумно уменьшить количество текста, передаваемого модулю ответа на вопрос» (Захаров 2015: 5).

Для русского языка существуют BOC Exactus и Deep Pavlov. В Exactus поиск ответов на вопрос пользователя осуществляется на основе выдачи поисковых систем (Google, Яндекс, Bing, Yahoo). Необходимо ввести вопрос на естественном языке в том виде, в каком его можно задать собеседнику, или описать ситуацию. Полученные результаты анализируются посредством

лингвистических инструментов Exactus и наболее релевантные документы выдаются пользователю» (Exactus).

DeepPavlov – это библиотека с открытым исходным кодом, предназначенная для разработки готовых чат-ботов и сложных разговорных систем, НЛП и исследования диалоговых систем. На сайте Deep Pavlov можно задать как свой, так и предложенный вопрос к тексту (Text QA). Для анализа статьи из Википедии, используется разметка страницы, где прописана основная информация и сам материал чётко структурирован. «Решение задачи идет в два шага: сначала подбираются релевантные документы, затем в тексте каждого выбирается фраза, предположительно содержащая ответ, и наиболее подходящий отображается на экране» (DeepPavlov).

Т.С. Черноморова и С.П. Воробьёв предлагают классифицировать все ВОС в рамках следующих изменений (см. рисунок 4):

- типы поддерживаемых вопросов – на какие вопросы система может давать ответы;
- типы поддерживаемых ответов – какого рода ответы может давать система;
- техника вывода вопроса или ответа по источнику информации (извлечение пассажей с ответами из текстов, извлечение фактов или генерация ответов на естественном языке);
- источник информации – структурированная или неструктурированная информация;
- домен знаний – открытая область знаний;
- направление – кто задаёт вопрос: пользователь или система.

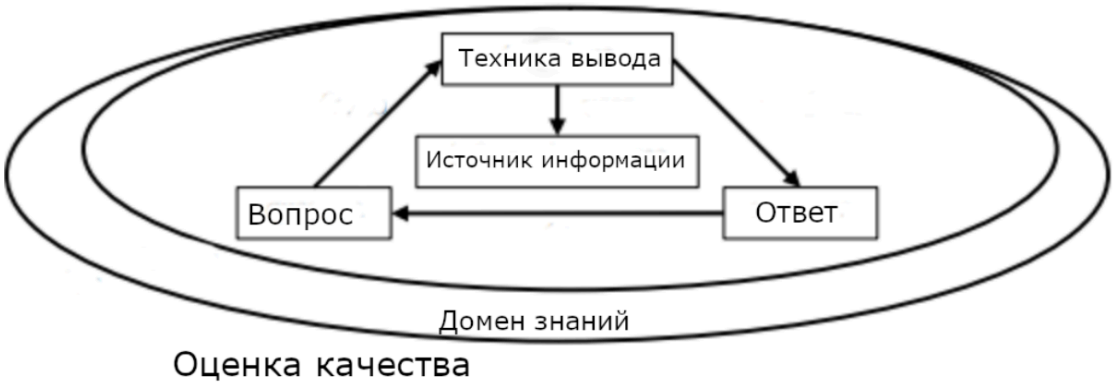

Рисунок 4 – Концептуальный фреймворк классификации (Воробьёв 2020: 151)

Данная модель охватывает главные аспекты в работе ВОС: вопрос и ответ, техника вывода, источник информации, а предложенная классификация ВОС несёт практически ориентированный характер.

Зарубежные исследователи также вносят свой значительный вклад в развитие вопросно-ответных систем. Например, авторы Paulo Cavalin, Flavio Figueiredo и др. разработали ВОС, предназначенную для выдачи ответов и обработки пользовательских вопросов, связанных с экономикой Бразилии. Источником информации послужили короткие предложения из социальных сетей специалистов в области экономики. Как считают авторы, достоинством коротких высказываний до 140 символов экономистов в Twitter является доступное, лёгкое для понимания, объяснение, надежность информации и отсутствие спама. Принцип работы этой ВОС следующий: собирается список аккаунтов экспертов по данной теме и сохраняются все их сообщения в базе данных; после чего модуль поиска вопросов находит все твиты, содержащие вопрос; затем модуль поиска ответа извлекает ответ из **ссылок,** которые были упомянуты в ответе (Paulo Cavalin 2016).

Нестандартное применение ВОС предложили исследователи Kemachart Kemavuthanon и Osamu Uchida из Университета Токай. Они разработали ВОС на случай стихийного бедствия для людей, проживающих в Японии, но не владеющих японским языком. Система должна подсказывать пользователям, как вести себя в чрезвычайной ситуации. QA-система так же, как и предыдущая, основывается на данных Twitter (более девяти миллионов твитов), собранных во время землетрясения в Осаке 18 июня 2018 года. Для

решения поставленной задачи все вопросы были классифицированы по определённым категориям, для поиска ответов использовалась онтология, применялись статистические методы для вычисления меры близости между вопросами. Результат исследования показал, что шаблоны типичных вопросов не подходят для практического развития, т.к. они порождают множество других вопросов и подвержены ошибкам. Авторы считают необходимым разработать интерфейс для ввода вопроса с «кнопками» быстрого доступа с самыми важными вопросами (Kemachart Kemavuthanon 2020).

Как известно, задача информационных и вопросно-ответных систем состоит не только в том, чтобы вывести ответ из базы знаний, он должен быть «осмысленным», т.е. включать ровно ту информацию, которая необходима для раскрытия вопроса. Исследователи Shivani G. Aithal, Abishek B. Rao и Sanjay Singh сделали попытку решить данную проблему при помощи модуля подобия вопросов. Прежде чем проанализированный вопрос будет задан ВОС, он будет сравниваться со сгенерированными вопросами к тексту, после чего будет производиться оценка близости между пользовательским вопросом и сгенерированным. Механизм схожести вопросов эффективно определяет вопросы, на которые нет ответа, и вопросы, не соответствующие теме пользовательского вопроса. Этот модуль помогает ВОС сконцентрироваться на решении только тех вопросов, на которые есть правильный ответ (Shivani G. Aithal 2021).

Как показал обзор примеров реализации русскоязычных и англоязычных ВОС, работа в данном направлении ведётся совершенно с разных точек зрения и для достижения разных целей. Тем не менее, ещё остался комплекс нерешённых проблем как для ВОС на английском языке, так и на русском.

Тенденции развития ВОС. Сегодня акцент в ВОС сместился с пополнения базы знаний в сторону задачи поиска правильного ответа, уже существующего в массиве данных. Этим массивом стал весь интернет, а поисковик – системой вывода. «Новый класс систем, цифровые помощники

которых активно развиваются в последнее время, сочетает в себе гибридные функции. Основа взаимодействия с помощником – это ввод на естественном языке, который классифицируется в зависимости от содержимого или в ВОС задачу («Сири, какова высота Эвереста»), или в задачу управления устройством при помощи команд естественного языка («Сири, поставь таймер на 15 минут»)» (Воробьёв 2020: 150).

В свете пандемии возникла проблема доступности образования для всех без очного посещения учебных заведений. Поэтому перед многими образовательными порталами возникла задача оцифровки академического материала, создания на их основе электронных курсов и осуществления дистанционного контроля, как одной из важнейших функций обучения. Для решения последней задачи стали активно применяться вопросно-ответные системы. Например, по прохождении одного блока курса учащимся предлагается ответить на открытый вопрос или пройти тест (промежуточный контроль). Таким образом, уже система задаёт вопрос пользователю и подтверждает/опровергает ответ, а пользователь этот ответ вводит. Кроме методической задачи по грамотному составлению контрольных заданий, существует проблема анализа ответов обучающегося, которая заключается в раскрытии определённого смысла с учётом ограничения объема лексем в ответе.

Выводы

Как показал анализ литературных источников, ВОС сегодня является актуальной областью компьютерной лингвистики. Задача ВОС состоит в том, чтобы правильно распознать вопрос и дать релевантный ответ. Вопросно-ответные системы делятся на *узкоспециализированные*, которые применяются в конкретных областях, и *общие*, которые осуществляют поиск ответов в любых областях знаний. QA-система состоит из пяти блоков: вопрос, анализатор вопроса, методы обработки естественных языков (NLP), генератор ответа, ответ.

Наиболее известными QA-системами для русского языка являются Deep Pavlov и Exactus. Первая библиотека предназначена для разработки чат-ботов и диалоговых систем. На сайте Deep Pavlov можно задать как свой, так и предложенный вопрос к тексту (Text QA). Exactus позволяет искать ответ на вопрос или описание ситуации по сети (Google, Яндекс, Yahoo, КиберЛенинка, Википедия). Ответ он не «формулирует», только выдаёт отрывки с ключевыми словами и ссылками на источники.

Основной трудностью для QA-системы является анализ вопроса. От корректности распознавания вопроса зависит точность выдачи ответа. Когда мы имеем дело с интервью, эта задача осложняется рядом факторов. Во-первых, интервью представляет собой диалог и состоит из диалогических единств *стимул – реакция,* которые трудно разделить (в результате ответ из базы знаний может лишь частично совпадать с вопросом пользователя). Во-вторых, первичная форма интервью – устная; как следствие, происходит влияние устной речи и коммуникативной ситуации: переспросы, инверсии, большое количество разговорной лексики, в случае с региональной прессой – лексика с национальным компонентом, слова на нерусском языке. Наконец, одну и ту же информацию можно запросить разными способами. В результате мы сталкиваемся с вариативностью классификаций вопросов и незакреплённым порядком слов в русском языке.

Таким образом, можно резюмировать, что точность ответа в ВОС во многом зависит от формулировки и анализа вопроса. Если для английского языка задача по извлечению фокуса и опоры вопроса ещё осуществима, то для русского языка необходим анализ почти на всех языковых уровнях и с применением дополнительных методов. Вышеперечисленные факторы необходимо учитывать при разработке вопросно-ответных систем.

Глава 2. Создание вопросно-ответной системы для электронного регионального издания

2.1. Специфика языка региональной прессы Якутии

Прежде чем мы перейдём к разбору механики нашей ВОС, рассмотрим лингвистические особенности региональной прессы Якутии на примере материалов ЯСИА.

Смена формата СМИ с традиционного на электронный значительно не изменила язык публицистики. Текст СМИ по-прежнему выполняет две главные функции: информирование и воздействие. В связи с чем язык публицистики изучается в двух аспектах. С точки прикладной лингвистики: речевое воздействие (стратегии, тактики и приёмы), проблема соотношения реальности и текста как её возможной интерпретации. И с точки зрения стилистики текста: специфика публицистического текста, активные процессы русского языка.

«Содержание публицистики, составляют не любые идеи, но идеи, имеющие социальный статус» (Солганик 2017: 5). События и факты, о которых рассказывает публицистика, разворачиваются в социальном пространстве. К примеру, в статье, посвященной преступности, социальным пространством может выступать семья (неблагополучные семьи), школа (нравственное воспитание) и др.

По единодушному мнению исследователей, главной тенденцией языка публицистики стала его демократизация. СМИ быстро реагируют на изменения в языке, «впитывая» неологизмы, разговорную и сниженную лексику, заимствования и явления из соцсетей. В последнее время чрезвычайно расширился круг существительных для обозначения лиц по роду деятельности (феминитивы). Раньше для их обозначения предпочтение отдавалось существительным мужского рода.

Также под влияние разговорного стиля подпал и синтаксис, «главным результатом которого оказался отход от «классических», синтагматически выверенных синтаксических конструкций, с открыто выраженными

подчинительными связями и относительной законченностью грамматической структуры» (Валгина 2003: 204). Все более активизируется и отчасти приходит на смену актуализированный синтаксис – выдвижение семантически значимых компонентов в актуальные позиции предложения с нарушением синтагматических цепочек (Валгина 2003: 204). В первую очередь это заметно в заголовках, которые строятся по схеме: название общей проблемы и конкретизирующие частные аспекты, детали или название места и события. Например: *«Крым: цветущий миндаль на фоне взрывов», «Фильмы, фильмы: алгебра и гармония».*

СМИ в значительной мере формируют языковое сознание и языковые вкусы общества. Журналисты стремятся подать материал небанально, для чего используют различные механизмы креативизации. И если раньше под креативностью понимали прежде всего языковую игру и изменение состава фразеологизмов, то сегодня в понятие креативный текст входит другое:

1. иронично-шутливый тон высказывания;
2. стилизация;
3. стилевая специфика («современный язык», «без сложных слов», «тропы, яркие сравнения», смешение элементов разных функциональных стилей);
4. композиционные средства (прецедентные тексты и заголовки);
5. трансформация привычных логических отношений в тексте (сопоставление разнородных понятий, «странный выбор предмета речи» (Панова 2017: 45).

Языковая игра сегодня воспринимается читателями как пример плохого вкуса, юмор времён телевизионных шоу «Кривого зеркала» и «Смехопанорамы». Современный публицистический текст должен соответствовать духу времени, быть написан на доступном языке и включать всё больше визуальных элементов. В журналистской практике требования к качеству новостного материала характеризуются триадой: ясность – краткость – яркость (Сабитова 2013: 23).

Специфика этнических или региональных СМИ заключается в тематике и языке публикуемого материала. Приоритетной темой в работе журналистов является народный уклад жизни местного населения. Основная тематика региональных изданий состоит в освещении социальных проблем, в информировании об изменениях в местном законодательстве и в знакомстве с выдающимися жителями города через интервью.

Республика Саха (Якутия) – это самый большой субъект Российской Федерации, на котором проживают представители более 120 национальностей: якуты, русские, украинцы, узбеки, китайцы и др. В связи с чем возникает вопрос о межъязыковом сосуществовании в регионе двух государственных языков – русского и якутского. В Якутии материалы СМИ публикуются как на русском языке, так и на якутском. Электронные издания предлагают пользователю возможность переключить язык. В традиционной форме выпускаются газеты и журналы на русском или якутском языках. Несмотря на это, происходит взаимовлияние языков: в речь якутов проникают русские иностранные слова и наоборот. Поэтому даже в газетах на русском языке есть якутские слова.

Таким образом, наиболее ярко национальная специфика представлена на лексическом уровне. ***Лексика с национально-культурным компонентом значения*** (далее ЛНК) – наиболее выразительная особенность языка региональной прессы. Она называет предметы и явления в жизни, которые отражают быт определенного народа и обладают особым культурологическим значением. Впервые систематическое исследование ЛНК было представлено в работе Е.М. Верещагина и В.Г. Костомарова, в разработанной ими на материале русского языка лингвострановедческой теории слова (Верещагин, Костомаров 1990).

В ЛНК выделяют традиционно четыре слоя: безэквивалентную, коннотативную, фоновую лексику и ключевые слова. Однако под национально-маркированной лексикой также понимают регионализмы – это слова и выражения, обозначающие реалии определённой местности и

функционирующие в устных и письменных текстах (Соколянская 2006: 27-28). Отличие регионализмов от ЛНК заключается в территории распространения: первая группа лексики, как следует из названия, будет понятна жителям региона, а вторая – жителям всей страны.

В прессе Якутии без труда можно найти примеры ЛНК и регионализмов, например, в заголовках: *«Якутские Игры Манчаары презентовали в Москве»* или *«Первомай в Якутске в прошлые годы. Большое собрание архивных фотографий ЯСИА»*. Первый заголовок содержит в себе незнакомое слово *Манчаары*, которое отсылает нас к имени якутского национального героя, борца против местных феодалов. Его именем назвали соревнования по национальным видам спорта, которые проводятся в Якутии раз в четыре года.

Второе предложение будет понятно жителям России, но не иностранцам. Скорее всего, название праздника вызовет затруднения и потребует семантизации. Первое мая – день весны и труда. Традиционно первого мая по всей стране проходят демонстрации трудящихся под лозунгом *«Мир! Труд! Май!»* Из объема и специфики знаний, а также территории распространения лексической единицы, мы можем констатировать, что словосочетание *Игры Манчаары* – регионализм, понятный только жителям Якутии, а слово *Первомай* – это ЛНК, знакомый всем россиянам. Такая экзотическая лексика никак не объясняется в тексте, т.к. предполагается, что читатель живёт в республике и знаком с национально-маркированной лексикой.

В общем составе ЛНК делится на следующие группы: топонимы, названия, связанные с духовной культурой (обряды, религия, спорт), с художественной культурой (декоративно-прикладное искусство, праздники), материальной культурой (архитектурные строения, предметы быта); прецедентные имена (имена и фамилии первых лиц республики, народных мастеров), этнонимы, фауномическая лексика.

Особый пласт лексики составляют якутизмы – слова из якутского языка, ассимилированные языком-рецептором (русским) или адаптированные на русский язык в результате влияния русской словообразовательной системы.

«Признаками ассимиляции традиционно признаются фонетическое, грамматическое освоение слова, словообразовательная активность, частотность употребления, семантическое освоение лексической единицы» (Кохан 2007: 1).

В составе якутского алфавита есть буквы из латиницы и кириллицы, поэтому на письме якутизмы не так просто распознать: они могут быть самостоятельным якутским словом, так и быть транслитерацией на русский. Например, слово *бар* в русском языке означает небольшой ресторан со спиртными напитками и закусками, но это же слово на якутском означает глагол *идти, ходить, ехать.* Или другой пример – одно из самых частотных слов в прессе Якутии – прилагательное *якутский* является транслитерацией слова на якутском языке – *дьокуускáй*. Как мы видим, при ассимиляции слово может сохранять своё значение, но претерпевать изменения в форме и наоборот.

В качестве примера проанализируем семантику лексемы с национально-культурным компонентом – *якутские алмазы.*

Якутские алмазы. В речевом обиходе можно встретить два словосочетания – *якутские бриллианты* и *якутские алмазы*. Они не являются синонимами: результатом обработки алмаза является бриллиант. В «Активном словаре русского языка» мы находим следующее определение слова *алмаз* и его лингвистические характеристики (см. рисунок 5). Что примечательно, в качестве примера неоднократно используется словосочетание *якутские алмазы* (Активный словарь русского языка 2014: 71):

АЛМА́З, СУЩ; МУЖСК; *-а.*
алмаз 1.1, обычно в форме МН.
Запасы алмазов; *найти в Якутии алмазы.*
ЗНАЧЕНИЕ. ‘Самый твердый минерал, являющийся кристаллической разновидностью углерода, обладающий большой промышленной и ювелирной ценностью’.
1. Алмазы относятся к разряду *драгоценных камней* и входят, наряду с *рубином, сапфиром* и *изумрудом*, в их первый класс. Основные разновидности алмазов – *баллас, карбонадо* и *борт*. Кристаллы алмаза могут быть окрашены в *черный, желтый, коричневый, красный, пурпуровый* и *голубой* цвета или быть *бесцветными*.
2. Алмазы часто встречаются в *кимберлитах*, или *кимберлитовых трубках*, называемых также *алмазными трубками*, – сверхглубоких природных скважинах, образовавшихся при прорыве газов сквозь земную кору: *«Удачная» – крупнейшая кимберлитовая трубка и месторождение алмазов на севере Якутии в России.*
3. Часто используется в названиях учреждений: *торговый центр <концерн> «Алмаз», НПО «Алмаз» имени академика Расплетина, центральное конструкторское бюро «Алмаз» в Петербурге.*
СОЧЕТАЕМОСТЬ. *Крупные <отборные> алмазы, сырые <необработанные> алмазы*; *природные алмазы, искусственные <синтетические> алмазы*; *алмазы ювелирного качества*; *месторождение алмазов, синтез алмазов*; *кристаллы алмаза*; *химический состав алмаза, физические свойства алмазов, вес алмаза, чистота алмаза*; *добыча <огранка, обработка> алмазов*; *рынок алмазов*; *искать <добывать> алмазы, выращивать алмазы, обрабатывать <шлифовать> алмазы.*

Рисунок 5 – Словарная статья «Алмаз»

Слово *алмаз* мы найдём в англ. яз. – *diamond,* в нем. яз. – *Der Diamant,* т.е. семантика слова содержит межъязыковое лексическое понятие и эквиваленты в других языках. Однако в сочетании *якутские алмазы* проявляется фоновая семантика, актуализируются национально-специфичные семантические доли.

В «Русском ассоциативном словаре» к слову-стимулу *алмаз* выдаются следующие реакции (список выстроен по мере убывания частоты): *камень, глаз, бриллиант, твёрдый, дорогой, гранёный, кольцо, блеск, блестит, богатство, гранит, драгоценный камень, изумруд, красота* (Русский ассоциативный словарь 2002). Таким образом, можно выделить следующие семантические доли, относящиеся к понятию *алмаз:* 1) твёрдая горная порода *(камень, твёрдый, драгоценный камень, гранит)*, 2) внешний вид *(блестит, блеск, красота, гранёный)*, 3) атрибут статуса *(дорогой, богатство),* 4) украшение, куда традиционно помещают бриллиант – *кольцо*. В ответах присутствует выражение *глаз алмаз (разг. шут.)* – умение увидеть главное, очень хорошее зрение. Также есть ассоциат *изумруд*: часто можно услышать сочетание *алмазы-изумруды* как самые дорогие драгоценные камни. По

данным ассоциативного словаря, среди ответов реципиентов не было обнаружено словосочетания *якутские алмазы.*

Считается, что происхождение бриллианта невозможно определить ни одним из существующих методов. Однако в рассматриваемом сочетании прилагательное указывает на место происхождения: *якутские* алмазы добываются в условиях вечной мерзлоты, чем объясняется высокая цена на изделие. Существуют также *канадские алмазы, африканские алмазы, смоленские бриллианты*. Если место добычи никак не определить, указание на место происхождения служит средством привлечения покупателей (брендом). В примере ниже (1) *якутские алмазы* включены в контекст, в котором перечисляются достопримечательности Дальнего Востока, что указывает на актуальный компонент фоновой семантики этого словосочетания.

> ***(1)*** *Так, 83% пользователей хотели бы побывать на Дальнем Востоке. Макрорегион, в основном, ассоциируется у них с амурскими тиграми, красной икрой, Курилами, Тихим океаном, камчатскими вулканами. В числе менее популярных ответов – северные олени (11% опрошенных), «Дальневосточный гектар» (11%), якутские алмазы (11%), Транссибирская магистраль (10%), Полюс холода в Оймяконе (4%).*
> ***(2)*** *Якутские бриллианты будут пользоваться спросом всегда. Это связано с тем, что количество населения планеты растет, и спрос будет только увеличиваться.*
> ***(3) — Что именно из набора представлений о Якутии (холод, алмазы, шаманы), на Ваш взгляд, может быть наиболее интересно миру?***
> ***(4)*** *Цель экспозиции – только дать общее понимание, сколько же богатств таят в себе недра Якутии, где одних алмазных месторождений 47.*
> *Алмазы до того момента, пока не превратятся в сияющие бриллианты, выглядят достаточно скромно. Если, конечно, минерал не потрясает своими размерами.*

Как показывает анализ контекстов, словосочетание *якутские алмазы* чаще всего упоминается в текстах рекламного характера (нативная реклама) и текстах, посвящённых туризму. Таким образом, фоновая семантика сочетания *якутские алмазы* в качестве важнейшего компонента включает следующие СД – достояние Якутии, достопримечательность, символ богатства недр республики, «бренд» в торговой сфере, марка на мировом рынке. Это не просто драгоценный камень, это и символ холода, достояние Якутии, которое может быть интересно туристам.

Локальная культура – это соединение норм, ценностей, символов, стереотипов, языка, восприятия народа. В локальной культуре коренных

народов Севера природа несёт почти божественный характер: с одной стороны, она кормит, одевает, защищает, но в то же время может и наказать человека за его неподобающее поведение. Например, в Якутии известен поселок Маган – монголизм, обозначающий белый цвет. Белый цвет – олицетворение чистого, священного, возвышенного. «Монголы и тюркские народы Сибири в особо торжественных случаях приносили в жертву богам животных белой масти. Их богатыри ездили на белых конях. <…> Село Крестях, Крест-Хальджай – «с крестом», «холм с крестом» – ассимилировавшееся русское слово «крест» (Прибылых 2019: 41).

«Язык для индивидуума может являться не просто инструментом общения, но и символом и средством, связывающим его с этническим коллективом, который и дает ему ощущение культурной и этнической идентичности» (Васильева 2013: 21). Якутские этносоциологи контатируют о некоторой языковой напряжённостью между двух наций (Васильева 2013: 28). Эта напряжённость, на наш взгляд, проявляется в контрасте культур и конкуренции за право лидерства между двумя нациями: якуты как коренной народ севера и русские как представители РФ, в состав которой входит Якутия.

2.2. Анализ поисковых систем СМИ Якутии

Пресса Якутии постепенно осуществляет переход из печатной, традиционной, формы в электронную. Причиной этого является постепенное снижение тиражей печатной продукции и эпидемиологическая обстановка в стране. Теперь издательства всё чаще осуществляют свою деятельность в digital-пространстве. В первую очередь они предлагают читателю электронные версии изданий и модернизируют сайты для большего удобства пользования. Электронная форма также позволяет не прочитывать статью до конца, а осуществлять поиск ключевых слов по странице или прямо задавать вопрос к тексту, т.е. использовать вопросно-ответные системы.

На сегодняшний день самые известные интернет-издания республики не имеют возможности прямо задавать вопрос к тексту или сайту (базе знаний).

Например, на главном новостном сайте Якутии *ykt.ru* размещены форумы, блоги, информационные сайты, публикации. Как пишут о себе редакторы сайта: «Ykt – региональный digital-media №1 в России по охвату населения в своем регионе» (Ykt.ru). В сутки его читают более 170 тыс посетителей. На сайте можно не только узнавать и самим публиковать новости, но и подавать бесплатные объявления, вести блоги, искать вакансии по всей республике. По всему этому многообразию возможно ориентироваться только при помощи разделов и поиска по сайту: при этом учитываются все формы слова, автор, дата и периодизация. Несмотря на то, что большое количество заголовков форумов и статей имеют вопросную структуру, задавать вопрос сайту всё ещё нельзя (см. рисунок 6). Хотя это сделало бы новостной сайт ещё более технологичным и удобным.

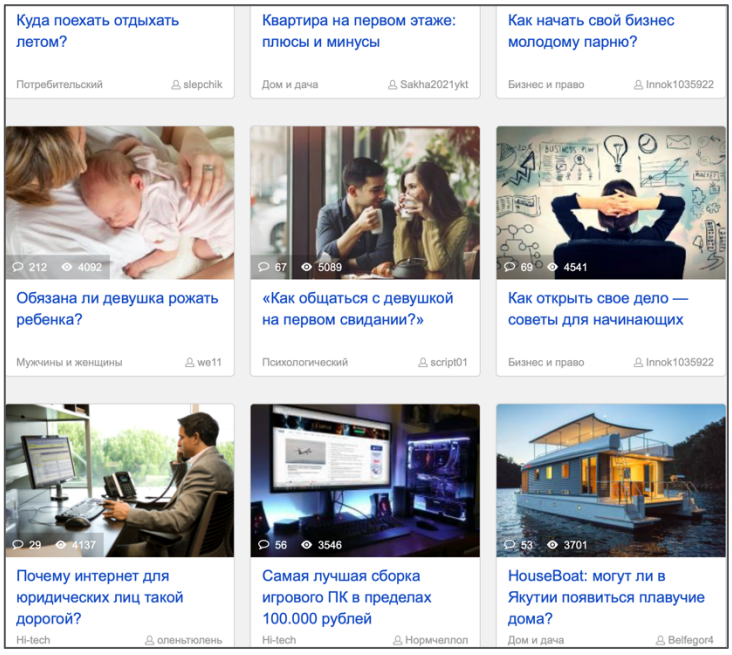

Рисунок 6 – Заголовки форумов с сайта Ykt.ru

Другая известная газета республики – **«Якутск вечерний»** – имеет печатную и электронную формы выпуска (vecherniy.com). В издании публикуются в основном криминальные новости и городские слухи. Печатная версия газеты формально считается самой тиражируемой – 60 тыс экз. Однако у газеты сформировалась негативная репутация из-за публикации недостоверных или преувеличенных фактов. Поэтому целевая аудитория издания – это непривередливый в достоверности получаемой информации

читатель, которому интересен развлекательный контент (читатели «желтой прессы»).

Поиск по сайту осуществляется по конкретной словоформе, а не по лемме, и только по заголовкам. Не учитывается автор, время публикации, период, рубрика, тэги. Например, если ввести в поисковую строку слово *погода*, то система не покажет нам ни одной записи, но если то же слово напечатать в предложном падеже – *погоде*, уже будет результат: *Мэр Якутска об аномальной погоде* (см. рисунок 7).

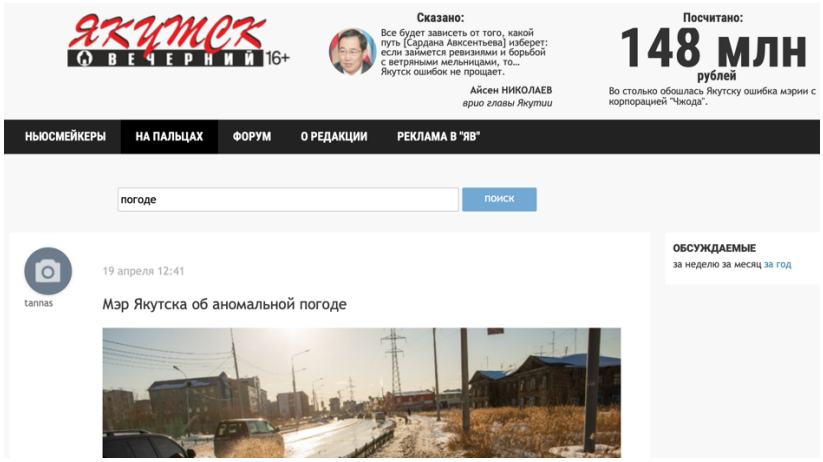

Рисунок 7 – Запрос в поисковой системе сайта «Якутск вечерний»

В тематическую группу погода не входят материалы, посвящённые ни вечной мерзлоте, ни резкому изменению температуры. В результате огромное количество материала оказывается вне поля зрения читателей, если автор не обозначил ключевое слово в заголовке. Получается, поисковая система на сайте «Якутск вечерний» не выполняет свою функцию – поиск информации по запросу.

Теперь обратимся к сайту ЯСИА (Якутское-Саха информационное агентство). Внешнее оформление и содержание электронной газеты отражают цель издания – информирование граждан, создание эффективных каналов связи государства и общества (см. рисунок 8).

Рисунок 8 – Сайт интернет-издания ЯСИА

Дискуссионный сайт ЯСИА максимально приближен к популярным в интернете блоговым платформам, что обеспечивает доверительный эффект у читателей, демонстрирует открытость власти. Обычно в ЯСИА публикуются новостные заметки и интервью, редко – лонгриды. Лонгрид (англ. *long read* – долгое чтение) – длинная статья, разбитая на части различными мультимедийными элементами (фотографиями, инфографикой).

В отличие от газеты «Якутск вечерний» на сайте ЯСИА поиск работает как по заголовкам, так и по основному тексту. Можно ограничить период *за всё время / за день / за неделю / за месяц / за год / за период* и рубрику – *актуально, в мире, видео, выбор редакции, здоровье, инфографика, образование, экономика, экология* и др. Система позволяет искать не только слова, но и словосочетания. Например, на запрос *цветок Сардаана* выдаются статьи, в которых есть оба слова из словосочетания: *День сардааны: Директор Ботанического сада СВФУ рассказала о символе Якутии*, *Фото: Цветок Солнца сардаана глазами Галины Давыдовой, В Хангалассском улусе расцвела сардаана с 36 бутонами.* Среди результатов оказалось две статьи, посвящённые кино в республики. Они оказались в выдаче из-за двух слов в тексте – ЦВЕТОКоррекция и женское имя Сардаана.

Таким образом, результаты поисковой системы сайта ЯСИА можно назвать релевантными: большая часть из них соответствуют запросу, поиск

осуществляется по лемме и внутри всего текста статьи, есть возможность фильтрации выдачи. Однако в поисковой системе есть несколько недостатков: отсутствует электронная клавиатура и горячие клавиши с буквами якутского алфавита, как это сделано на портале *Ykt.ru*, не учитывается омонимия. Например, если мы введём название реки Лена, то предсказуемо среди результатов будет много статей, в которых есть женское имя Лена или Елена. То же самое будет, если мы введём слово *Сардаана* без конкретизации, что имеем в виду цветок.

Для первого правительственного издания, освещающего новости по всей Якутии, наличие ВОС расширило бы доступ к проверенной информации, и в перспективе повысило бы конкурентоспособность на рынке якутских СМИ. Пользователи могли бы не обращаться к частным, малодостоверным, изданиям, а задавали бы интересующие вопросы о республике на сайте ЯСИА.

2.3. Модель и механизм работы вопросно-ответной системы

Модель ВОС, которую мы предлагаем для регионального интернет-издания ЯСИА, состоит из 1) вопроса пользователя, 2) базы знаний (собранные вопросы и ответы), 3) векторной модели, которая трансформирует существующие вопросы в базе знаний в векторы, 4) анализатора вопроса пользователя (токенизация, лемматизация), 5) векторной модели для пользовательского вопроса, 6) поиска ответа исходя из сходства косинусной меры входящего вопроса и существующих вопросов в базе, 7) выдачи наиболее релевантного ответа. Модель работы ВОС для региональной прессы Якутии изображена на рисунке 9:

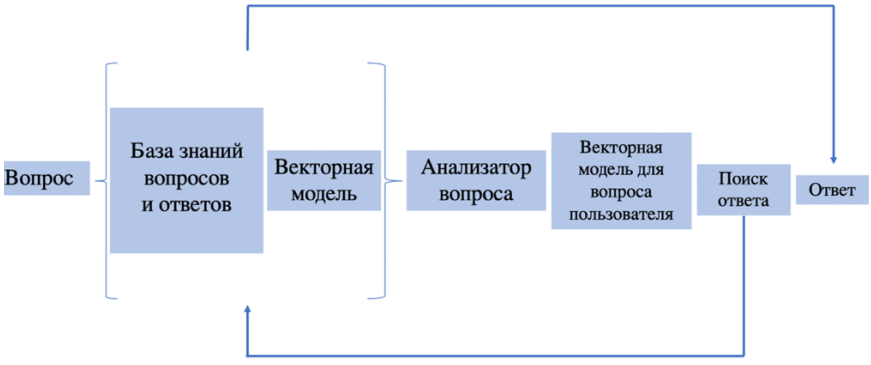

Рисунок 9 – Модель ВОС для регионального интернет-издания ЯСИА

Источником информации являются тексты интервью из электронного издания ЯСИА. Вопросы в интервью выделялись с применением регулярных выражений, которые реализованы в пакете re. Интервью представляет собой совокупность изречений интервьюера и интервьюируемого, при этом на письме они часто оформляются в виде прямой речи, которая начинается с тире. Поэтому формальная запись вопроса представлена нами как **[0-9А-Яёа-яёһҕөүҥ\-].** Нами было замечено, что в тексте иногда используются слова на якутском языке, поэтому в шаблон регулярного выражения мы включили буквы якутского алфавита.

Количество вопросов и ответов в базе знаний составляет 1303 вопроса и 1303 ответа, собранных за 2019 и 2020 год. В базе знаний находятся вопросы разных типов. Например, есть открытые вопросы на тему здравоохранения: *Как сегодня обстоит дело с районными больницами? Многие роддома закрываются, что делать беременным?* – и закрытые вопросы: *Надо ли принимать витамины?* Также есть количественные вопросы: *Как долго региональные каналы останутся в аналоговом вещании? Каков охват населения и насколько приблизились к отметке систематически занимающихся физкультурой и спортом в 55 процентов от общего числа населения?* В тематическом плане вопросы самые разнообразные: экономика, развитие Дальнего Востока, школьное и дошкольное образование, строительство моста через Лену, доступное жильё и другие волнующие проблемы населения.

Суть работы данной ВОС заключается в следующем: все 1303 вопроса мы представили в виде векторов, разделили их на общее количество и в результате создали векторное представление вопросов. Вопрос пользователя мы также представили в виде вектора. После чего вектор пользовательского вопроса сравнивается с векторами существующих вопросов и на основании косинусной меры выводится ответ.

Задача системы – получить и распознать вопрос пользователя и дать на него релевантный ответ из базы знаний.

Проанализируем подробно механизм работы программы (см. приложение А).

1. Импортируем необходимые библиотеки:
   - Pymorphy 2 – для морфологического анализа слов и приведения слов к начальной форме слов (иначе одно и то же слово в разных формах будет считываться как новое);
   - Openpyexcel – для открытия и считывания ячеек из Excel-файлов;
   - Gensim Word2vec – для представления всех вопросов в векторы;
   - Re – для шаблона извлечения вопросов;
   - Numpy – для работы с массивом данных (т.к. вектор – это набор случайно присвоенных чисел, массив);
   - модули dot и norm – для подсчёта косинусной близости.
2. Загружаем файл Excel со всеми вопросами и ответами, прописываем номер листа Excel, создаём два списка *questions* и *answers*.

```
questions = []
answers = []
```

3. Распределяем все вопросы и ответы из общего файла в списки. В переменных *A* и *B* мы будем хранить адреса самих ячеек, а в переменных *q* и *a* – значения ячеек.

```
for i in range(2, 1305):
  a_value = 'A'+str(i)
  b_value = 'B'+str(i)
  q = sheet[a_value].value
  a = sheet[b_value].value
  questions.append(q)
  answers.append(a)
```

4. Открываем список стоп-слов, удаляем переносы и записываем все стоп-слова в список *swlist*, чтобы в дальнейшем исключить их из списка лемм. Список стоп-слов был создан на основе словаря О.Н.

Ляшевской и С.А. Шарова (Ляшевская, Шаров 2009). Его объём составляет более 1400 частотных словоформ.

5. Создаём список *final_pr*, куда будем записывать токенизированные и лемматизированные вопросы, а в переменной *morph* мы будем записываем функцию для морфологического анализа. Из списка *questions* мы извлекаем все лексические единицы и записываем его в список *lst*, приводим к словарной форме (список *new_list*), проверяем их наличие в списке стоп-слов. При морфологическом анализе мы берём первое значение из OpenCorpora.

```
final_pr = []
morph = pymorphy2.MorphAnalyzer()
for i in questions:
   lst = re.findall(r'[0-9A-Яёа-яёһҙөүһҥ\-]+', i)
   new_list = []
   for w in lst:
    k = morph.parse(w)[0]
    k = k.normal_form
     if k not in swlist:
      new_list.append(k)
final_pr.append(new_list)
```

Представление векторной модели

6. Создаём функцию *model,* к которой применяем функцию Word2vec. На вход мы подаём токенизированные и лемматизированные вопросы. *Size window* – размер вектора, который улучшает качество обучаемости модели. Благодаря данному параметру, модель «научится» предсказывать, однако если выбрать большой размер, система может «переобучиться» и «запомнить» вопросы. *Window* – контекстное окно, которое оценивает четыре слова слева и четыре слова справа, чем задаёт характеристику главного слова (парадигматические связи). *Workers=4* означает максимальные возможности сервера. *Min_count=1* – минимальная частота вхождения. Все эти переменные позволяют не переобучить систему.

```
model = Word2vec(final_pr, size=300, window=4, min_count=1,
workers=4)
model.init_sims(replace=True)
model.save('qa_system_w2v.model')
```

7. До этого этапа все вопросы были представлены в виде большого списка списков. Теперь задача состоит в том, чтобы представить все вопросы в виде векторов. Для этого мы берём оттокенизированный, отлемматизированный и проверенный стоп-словарём вопрос (первый подсписок в общем списке) и переводим каждое слово в вектор (массив) с помощью Numpy и Word2vec. Всё это записываем в переменную *vector_means*. Если слово не найдётся (ошибка), мы его пропускаем при помощи функции *try – except.* Мы также вводим переменную *vector* – векторное представление слова и используем атрибут *model.wv*, который расшифровывается как *word vector*. В квадратных скобках мы прописываем, какое слово нужно перевести в вектор. В результате получается массив, записанный в переменную *vector.* Затем мы добавляем его в список *vetor_mean,* в котором хранятся векторные значения всех слов из вопроса.

```
vector_questions= []
for quest in final_pr:
  vectors_mean = []
  for y in quest:
    try:
      vector = model.wv[y]
      vectors_mean.append(vector)
    except:
      pass
```

8. Теперь вычисляем средний вектор BOC структуры: складываем все вектора и делим на длину *vector_mean* (количество векторов). Присоединяем векторную форму к общему списку для сравнения пользовательского вопроса с базой знаний.

```
try:
    average = sum(vectors_mean)/len(vectors_mean)
    vector_questions.append(average)
```

Действия, описанные в пункте 7 и 8, мы повторяем с каждым вопросом. Как мы знаем, вопросы по своей структуре бывают достаточно разными: открытые и закрытые, уточняющие, развивающие, контрольные и т.д. Это создаёт некоторую трудность для работы ВОС. Например, уточняющий вопрос *«Почему?»,* который обычно следует после закрытого вопроса, в результате обработки может оказаться в списке стоп-слов, т.к. данный вопрос будет идентифицирован как пустой список – ноль. Произойдёт ошибка деления. Поэтому, если возникнет ситуация, когда вопрос не будет распознан, мы создадим при помощи библиотеки Numpy массив нулей и запишем его в общий список векторов. *Vector_questions[-1]* – размер массива нулей должен совпадать с размером предыдущего вопроса.

```
except:
    average = np.zeros_like(vector_questions[-1])
    vector_questions.append(average)
```

Наша задача – создать систему, к которой можно задать вопрос, посвящённый Якутии и получить релевантный ответ. Мы вводим вопрос, токенизируем, лемматизируем, проверяем на список стоп-слов, переводим в векторную форму, создаём средний вектор, сравниваем средний вектор со всеми 1303 векторами, которые хранятся в базе знаний. В качестве ответа выбирается тот вектор, у которого косинусная мера будет ближе к единице.

9. Создаём функцию ввода и сравнения вопроса с теми векторными представлениями, которые есть в базе знаний, – qasystem.

```
def qasystem():
  user = input('Введите Ваш вопрос, посвящённый Якутии: ')
```

В токенизированном, лемматизированном и нормализованном виде записываем вопрос пользователя в список *new_list*. Проверяем на наличие стоп-слов.

```
tokens = re.findall(r'[0-9а-яёһөүҥ\-]+', user)
  new_list = []
  for token in tokens:
```

```
    k = morph.parse(token)[0]
    k = k.normal_form
    if k not in swlist:
      new_list.append(k)
```

10. Мы обращаемся к каждому слову из списка *new_list*, переводим каждое слово в модель (массив чисел) и записываем числовые данные в новый список *user_arr*.

```
user_arr = []
  for y in new_list:
    try:
      vector = model.wv[y]
      user_arr.append(vector)
    except:
      pass
```

11. Вычисляем средний вектор пользовательского вопроса: суммируем векторы всех слов в пользовательском вопросе и делим на длину списка. *Average_user* – усреднённый массив пользовательского вопроса. Если возникает ошибка, присваиваем нули.

```
try:
    average_user = sum(user_arr)/len(user_arr)
  except:
    average_user = np.zeros_like(vector_questions[-1])
```

12. Теперь нам необходимо сравнить *косинусную меру* сходства между пользовательским вопросом и каждым вопросом в нашей базе знаний, который хранится в переменной *vector_questions*. *Косинусная мера* – это мера сходства между двумя векторами, которая используется для измерения косинуса угла между ними (от 0 до 1). Для каждого усреднённого векторного вопроса в списке *vector_questions* находим косинусную близость с вопросом пользователя. Мы сравнили введённый пользовательский вопрос с каждым вопросом из нашей базы знаний, полученный показатель мы записали в *all_cosines*. В результате получается 1303 значения и столько же получается ответов.

```
all_cosines = []
  for i in vector_questions:
```

```
    cos_sim = dot(i,
average_user)/(norm(i)*norm(average_user))
    all_cosines.append(cos_sim)
```

Нам важно извлечь максимальное значение из списка all_cosines и его индекс, чтобы вывести ответ.

```
max_index = all_cosines.index(max(all_cosines))
```

13. Выводим ответ из базы знаний с наиболее высоким индексом.

```
final_a = 'Ответ из базы данных: '+answers[max_index]
return final_a
```

Результаты выдачи. Пример 1. Например, мы задаём системе следующий вопрос: *Как вы думаете, сельские жители не работают на предприятиях*? И получаем ответ из базы знаний: *«Не согласен с этим мнением. Людьми надо заниматься, их необходимо обучать, в том числе и на практике. Всем известная компания «Алмазы Анабара» на 80% состоит из представителей местного сельского населения. Они работали над этим вопросом долгое время и добились успеха. Предприятия постоянно должны уделять внимание своим работникам, адаптировать их».*

Тематически ответ соответствует нашему вопросу: система дала отрицательный ответ *«не согласен с этим мнением»* и привела иллюстрирующий пример, когда известная алмазодобывающая компания нанимает представителей сельского хозяйства. Сомнение вызывает первое предложение, где открытый вопрос был назван мнением. В связи с тем, что система не генерирует собственный ответ, а выдача представляет собой реальный ответ человека на какой-то другой вопрос, возникают неточности при подмене вопроса. В оригинальном диалоге журналист высказал мнение: «*Может сельские жители просто не подходят для работы на промышленных предприятиях?»*, на что получил ответ интервьюируемого «*не согласен с этим мнением».* В нашем же случае интенция заключалась в запросе информации, а не в выражении мнения.

Пример 2. Приведём другой пример. Мы ввели вопрос: «*Будет ли развиваться Заполярье?»* И система выдала следующий ответ: *«В последнее время наша культура начала осознавать свою азиатскую составляющую, несмотря на российско-европейское образование. Но сегодня следует весь фронт культуры развернуть к арктическому вектору, не только из-за российской политики».* Вопрос по структуре является закрытым, т.е. предполагает или положительный, или отрицательный ответ. На основании сходства векторов система выдала ответ, предполагающий открытый вопрос. Несмотря на это, мы считаем, что ответ в целом удовлетворяет запрос.

Пример 3. В связи с тем, что база знаний ограничена количеством документов, не все вопросы правильно распознаются системой. Так, на вопрос «*Какова ситуация с кино в Якутии?»* мы получили следующий ответ: *«Здесь всё зависит от наших совестных действий, от того, насколько мы будем соблюдать режим самоизоляции, насколько мы будем ответственными перед самими собой, близкими и окружающими. Я, конечно, не врач, но мы готовы к* ***любому сценарию*** *… Как я уже сказал, мы активно взаимодействуем с провинцией Хэйлунцзян. На прошлой неделе мы, по моей инициативе, организовали* ***телемост****».* Данный ответ совершенно не относится к вопросу. На основании тематически близких слов *(сценарий и телемост)* ответ из базы знаний оказался наиболее вероятным. Не учитывалось при этом переносное значение выражения *готовиться к любому сценарию* – быть готовым к неожиданному изменению ситуации.

Пример 4. Мы не могли не задать системе, самый волнующий вопрос для якутян, который неоднократно задавался на пресс-конференции президенту страны: *«Когда построят мост через Лену?».* Ответ из базы данных следующий: *«Во-первых, решение по мосту через Лену появилось лет десять назад. В 2008 году уже была подготовлена проектно-сметная документация совмещённого моста, которая прошла госэкспертизу. Потом начались долгие препирания: строить мост или тоннель. В итоге всё*

затянулось, а после появились другие государственные приоритеты». Мы считаем ответ из базы знаний полностью удовлетворяет вопрос.

Таким образом, ВОС осуществляет корректную выдачу ответов при условии схожести пользовательского вопроса с вопросом из базы знаний. Главным ограничением для пользовательского вопроса является то, что вопрос должен быть посвящен Республике Саха (Якутия).

2.4. Оценка результатов работы вопросно-ответной системы

В настоящем исследовании необходимо также оценить эффективность работы нашей ВОС. Для этого мы воспользуемся такими параметрами, как точность (precision), полнота (recall) и F-мера. Для их расчёта мы прибегнем к использованию таблицы сопряженности. В статистике таблицы сопряжённости чаще всего представлены в следующем виде:

Таблица 3 – Сопряжённость положительной и отрицательной оценки системы и эксперта

		Оценка эксперта	
		Положительная	Отрицательная
Оценка системы	Положительная	TP	FP
	Отрицательная	FN	FP

В таблице в рамках данной работы отражается информация о количестве верно и неверно выданных ответов на определённые пользовательские вопросы. Под TP мы понимаем истинно-положительные решения, под TN – истинно-отрицательные решения, под FP – ложно-положительные решения, FN – ложно-отрицательные решения.

На основании вышеприведённой таблицы в статистике выводятся следующие формулы для расчёта точности, полноты и F-меры.

1. Под точностью понимают частное от деления истинно-положительных решений на сумму истинно-положительных и ложно-положительных решений.

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

2. Под полнотой понимают частное от деления истинно-положительных решений на сумму истинно-положительных и ложно-отрицательных решений.

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

3. Под F-мерой понимают среднее гармоническое полноты и точности.

$$F - measure = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (3)$$

Для проведения экспертной оценки необходимо задать серию пользовательских вопросов системе. Мы ограничимся выборкой в 35 вопросов. В таблице 4 приведены некоторые примеры вопросов и полученные ответы.

Таблица 4 – Примеры пользовательских вопросов и выведенных ответов в ВОС

Пользовательский вопрос	**Вопрос-эталон**	**Выданный ответ**
Будет ли развиваться Заполярье?	Есть какая-то надежда поднять Заполярье?	До того уровня невозможно. Пытаются возродить Северный Морской путь, но комплексный закон по развитию Крайнего Севера не принимают. Кадры будут набирать вахтовым методом. Это для нас совсем не годится. Парламент и руководство Якутии ставят вопрос о комплексном законе по развитию Арктической зоны РФ.

Перейдёт ли издательство ЯСИА на удалёнку?	В связи с пандемией коронавируса многие перешли на «удаленку». В перспективе возможно применить такой формат работы в ЯСИА?	Пандемия продемонстрировала возможности Интернета и онлайн-технологий. Полностью, конечно, мы не перейдем на такой формат, но использовать можем, например, в работе с районами республики и другими регионами страны. А вся работа информагентства все равно будет проходить в оффлайн-формате. Живое общение иногда важнее и эффективнее.
Известна дата появления инвестиционного суда?	Когда на Дальнем Востоке появится инвестиционный суд?	Это сложный процесс, в том числе и для меня лично, потому что нельзя вносить руководству страны предложения, когда ты не можешь их аргументировать. Я прекрасно понимаю, какие вопросы мне зададут. Фрагменты ответов на эти вопросы существуют, потому что сопровождение экономических процессов, тем более с международными компаниями, предусматривает и другую специализацию, в том числе и судей. Но не все так просто. Мы обязательно расскажем, когда будет готовый продукт. Я уже не первый год работаю в правительстве и понимаю, в каких вопросах готов

		докладывать и надеяться на положительное решение, а в каких лучше еще подготовиться. Пока этой уверенности нет, бессмысленно выдвигать инициативу.
Опишите цели, которые оставил Вам глава Якутии.	Какие задачи перед Вами поставил глава Якутии?	Его коронное блюдо – плов, который он обязательно готовит на 8 Марта. В этот день мы расслабляемся в предвкушении папиного плова. По возможности он колдует над казаном и по другим семейным праздникам.
Что известно про якутское волонтёрское движение?	Как вы оцениваете развитие волонтерского движения в Якутии в настоящее время?	Назову три. Это, прежде всего, игры «Дети Азии», о которых я уже говорил. Отмечу только, что если в МСИ 2012 года было задействовано около тысячи волонтеров, то в 2016 году – уже в два раза больше. Кроме того, нельзя не вспомнить Олимпийские игры в Сочи, где, по словам президента страны Владимира Путина, добровольцам удалось создать особую атмосферу спортивного праздника. Третье – это решение объявить 2018 год в России Годом волонтерства. Думаю, что в условиях все более нарастающей коммерциализации

		особую ценность приобретают бескорыстные поступки, желание помочь окружающим.

Среди рассматриваемых примеров можно выделить следующую особенность системы, обученной на векторных представлениях. Модель дистрибутивной семантики позволяет учесть некоторые парадигматические связи между словами в задаваемых вопросах и вопросах-эталонах (например, синонимы: *«когда» – «известна дата»*, *«развивать» – «поднять»* и т.д.), что позволяет варьировать синтаксическую структуру самого вопроса с сохранением семантической составляющей. Конечно, даже при сохранении частичной синтаксической структуры вопроса система может дать сбой (см. таблицу 4, пример 4), что объясняется итоговыми параметрами обучения Word2vec модели.

Теперь обратимся к расчёту эффективности. Из 35 заданных вопросов система выдала 10 ошибочных ответов.

$$Precision = \frac{25}{25 + 10} = 0.71 \quad (4)$$

Под TP мы понимаем те вопросы, на которые были даны верные ответы, а под FP – ошибочные.

При расчёте полноты важно понимать, что под переменной FN скрываются те вопросы, на которые система не подобрала бы ответы. Поскольку в рамках разработанной системы ответы выводятся на основе высчитанной косинусной меры, то переменная FN всегда будет равна 0, так как система выведет наиболее близкий ответ, даже если он неверный.

$$Recall = \frac{35}{35 + 0} = 1 \quad (5)$$

F-мера для вопросно-ответной системы:

$$F - measure = 2 * \frac{0.71 * 1}{0.71 + 1} = 0.83 \quad (6)$$

Таким образом, полученные результаты свидетельствуют о том, что Word2vec модель может быть пригодна для разработки вопросно-ответных систем для региональной прессы.

Выводы

Современный ритм жизни диктует условия, при которых человек нуждается в быстром и удобном способе получения информации. Неторопливому чтению печатных газет и журналов на смену пришли электронные издания, которые можно читать «на ходу». Несмотря на то, что значительная часть СМИ в настоящее время уже адаптировалась к digital-пространству и осуществила переход в электронный формат, модернизация новостных сайтов на этом не заканчивается. Скорость, удобство, креативная подача материала, экономия времени – вот, вероятно, главные требования к федеральным и региональным периодическим изданиям.

Как показал анализ наиболее известных интернет-газет Якутии, республиканские СМИ уделяют большое внимание технологичности и дизайну своих сайтов. Например, развлекательная интернет-газета Ykt.ru предлагает пользователям не только читать новости, но и самим их публиковать, а также подавать бесплатные объявления, участвовать в обсуждениях на форумах, вести блоги, поэтому неудивительно, что Ykt читают более 170 тыс пользователей в сутки.

Другая общественно-политическая газета ЯСИА (***ysia.ru)*** преследует иную цель – освещать деятельность правительства и рассказывать о жизни местного населения. На сайте не так много развлекательного контента, как на Ykt.ru в силу целевого направления издания, тем не менее, как показал сравнительный анализ, ЯСИА обладает самой удобной поисковой системой.

Всё вышеизложенное подводит нас к выводу, что ВОС расширило бы возможности взаимодействия с электронными изданиями, повысило бы

конкурентоспособность в сфере региональных СМИ. Пользователю не приходилось бы прочитывать статью до конца, обращаться к другим СМИ (потенциальным конкурентам издательства), чтобы получить ответ на свой вопрос.

На основе модели Word2vec мы создали ВОС для региональной прессы. Механизм работы нашей ВОС близок к классической, но отличается модулем векторизации. Суть работы ВОС заключается в следующем:

1. Пользователь вводит вопрос, посвящённый Республике Саха (Якутия) *(например, где Якутия может найти партнёров?)*.
2. Предварительно мы собрали базу знаний материалов интервью из 1303 вопросов и 1303 ответов и создали их векторное представление.
3. Пользовательский вопрос представляется в виде вектора и сравнивается с векторами существующих вопросов.
4. Тот ответ, который оказался наиболее близким по косинусной мере с пользовательским вопросом и вопросом из базы знаний, выводится пользователю.

В целом система показала удовлетворительные результаты: из 35 вопросов 10 были ошибочными, точность составляет 0.71, полнота 1 (система всегда выдаёт ответ, правильный он или нет), F-мера – 0.73. Ошибки вызваны ограниченностью базы знаний и в некоторых случаях тесной взаимосвязью вопроса и ответа в интервью. Сделанные подсчёты приводят нас к главному выводу: вопросно-ответная система, обученная на векторных представлениях, может анализировать семантику пользовательского вопроса и может использоваться в региональных электронных изданиях, при учёте специфики местной лексики.

Заключение

Метод моделирования сегодня применяется во многих научных областях, и особенно там, где объект науки недоступен непосредственному наблюдению. Это одна из основных категорий теории познания: «на идее моделирования по существу базируется любой метод научного исследования – как теоретический (при котором исследуются различного рода знаковые, абстрактные модели), так и экспериментальный (использующий предметные модели)» (БЭС 2000). Цель моделирования состоит не только в замещении реального объекта действительности моделью, но и в выявлении некоторого нового знания об объекте, открытии содержимого «чёрного ящика» (Ю.Д. Апресян). «Рассвет» этого метода в лингвистике пришёлся на 60-70-е годы XX века и стал активно использоваться в математической лингвистике.

Так как речевая деятельность человека является сложнейшим актом, модели, её отражающие представляют собой сложную и разнообразную систему. Объектом моделирования в лингвистике может быть коммуникативная ситуация, языковая картина мира, структура предложения, организация текста и многое другое. В нашей работе моделирование рассматривается в аспекте создания вопросно-ответной системы для электронного регионального издания.

Вопросно-ответная система (от англ. Question Answering Systems) – это вид информационно-поисковых систем, способных обрабатывать введённый пользователем вопрос на естественном языке и выдавать осмысленный ответ. Архитектура типичной QA-системы состоит из пяти блоков: вопрос, анализатор вопроса, методы обработки естественных языков (NLP), генератор ответа, ответ.

Диалоговое взаимодействие пользователя с машиной протекает в одном из трёх режимов: 1) когда на вопросы системы отвечает пользователь, 2) когда на запрос пользователя определённым образом реагирует система, и, наконец, 3) двухсторонне активный диалог, когда пользователь и система меняются ролями в ходе общения (Сулейманов 2016: 27). Например, первый режим

активно используется в образовательной среде и онлайн-курсах; второй режим является наиболее изученным и применяется в различных сферах: в бизнесе, экономике; третий режим, наоборот, малоизучен, его примером может быть чат-бот, или виртуальные ассистенты, когда система пытается «вести диалог» (имитировать беседу), но чем больше она уточняет, тем больше этот диалог превращается в неразбериху. Наиболее известные ВОС – START, BASEBALL, ВОС Lasso, Exactus и DeepPavlov.

Как мы неоднократно упоминали, основная трудность в работе ВОС заключается в распознавании семантики вопроса (модуль обработки вопроса), которая зависит от темы диалога, коммуникативной ситуации и специфики выбранных языковых средств. Поиск ответа осуществляется в источнике информации: это может быть массив данных глобальной сети или ограниченная база знаний. Сегодня при разработке ВОС основной упор делается не на пополнении источника информации (им стал интернет), а на задачи распознавания вопроса и поиска правильного ответа, уже существующего в массиве данных – т.е. на достижении ВОС коммуникативной цели. В целом ВОС должна выполнять три главные задачи: анализировать семантику вопроса пользователя, осуществлять поиск ответа в источнике информации, выводить наиболее релевантный ответ.

Для извлечения сути вопроса на естественном языке используются различные методы. Большая часть из них так или иначе базируются на подходе, при котором выделяется фокус вопроса – вопросное словосочетание, опора вопроса – остальная часть вопроса после «вычета» фокуса, и присваивается семантический тэг будущего ответа, согласно некоторой ранее заданной таксономии (например, Date, Money, Yes/No).

Анализ вопроса производится при помощи регулярных выражений, когда задаётся символьный шаблон, например, в нашем исследовании – [0-9A-Яёа-яёhҕөүннҥ\-]; при помощи метода построения синтаксических деревьев, когда из вопросительного предложения вычленяются элементарные синтаксические связки, а в их узлах задаётся аннотация с морфологическими

метками; активно используются семантические тэги в вопросах и предполагаемых ответах и статистические методы (например, мера TF*IDF – для оценки важности слова в контексте документа).

Для достижения поставленной цели – создание прототипа русскоязычной ВОС для электронного республиканского издания ЯСИА – мы выявили специфику языка региональной прессы Республики Саха, провели анализ существующих поисковых систем в электронных СМИ республики, схематично представили модель будущей ВОС, написали программу и провели оценку результатов работы по таким параметрам, как точность (precision), полнота (recall) и F-мера. Имеющиеся в нашем распоряжении данные, позволяют нам сделать выводы по каждой предметной области.

Специфика этнических СМИ в отличие от федеральных заключается в выборе тем публикуемого материала и языке, на котором он пишется. Это может быть как язык титульной нации, так и русский язык с местными наименованиями. Лексика с национально-культурным компонентом – наиболее выразительная особенность региональной прессы. Она называет предметы и явления в жизни, которые отражают быт определенного народа, обладают особым культурологическим значением и которые не имеют эквивалентов в иностранных языках (например, *Игры Манчаары, цветок сардаана, летний праздник ЫСыах).*

В общем составе ЛНК делится на следующие группы: топонимы, названия, связанные с духовной культурой (обряды, религия, спорт), с художественной культурой (декоративно-прикладное искусство, праздники), материальной культурой (архитектурные строения, предметы быта); прецедентные имена (имена и фамилии первых лиц республики, народных мастеров), этнонимы, фауномическая лексика. Знание лингвистических особенностей региональной прессы необходимо и важно не только для анализа вопроса, в составе которого есть ЛНК, но и для обработки фрагмента текста, где содержится ответ, и для применения всех тех методов, которые мы перечислили выше.

Как показал анализ поисковых систем самых посещаемых интернет-изданий Якутии (Ykt.ru, vecherniy.com, ЯСИА), вопросно-ответные системы пока не получили широкого распространения в республике несмотря на то, что большое количество заголовков форумов и статей имеют вопросную структуру. В настоящее время поиск материала осуществляется в лучшем случае (поисковая система ЯСИА) по тексту и заголовкам, а в остальных случаях – только по заголовкам или рубрикам. Внедрение в республиканское издание ВОС позволило бы узнавать новости от официального источника, прямо задавая вопрос сайту (напомним, ЯСИА – правительственное издание), сделало бы сайт удобнее для пользователя, чем повысило бы конкурентоспособность на рынке региональных СМИ.

ВОС, которую мы разработали для региональной прессы Якутии, основывается на модели Word2vec, позволяющей анализировать семантику слов на естественном языке с помощью векторов. Архитектура нашей ВОС не сильно отличается от стандартной, за исключением модуля Word2vec, который представляет вопрос пользователя, а также все вопросы и ответы из базы знаний в виде векторов (1303 вопроса и 1303 ответа) и анализирует степень их близости с пользовательским вопросом. В качестве источника информации использовался корпус интервью, собранных с 2019 по 2020 год, с сайта ЯСИА. Также для работы нам потребовались библиотеки Pymorphy 2, Openpyexcel, Gensim Word2vec, Re и Numpy.

Работа системы показала, что из 35 заданных вопросов 10 были ошибочны. Таким образом, точность составила 0.71, полнота 1 (система всегда выдаёт наиболее близкий ответ), F-мера – 0.83. Основные ошибки вызваны 1) ограниченностью базы данных (на некоторые вопросы системе не хватало «ресурсов», чтобы ответить); 2) тематическим несоответствием пользовательского вопроса (система «на вход» принимает вопросы, связанные с Республикой Саха); 3) неверным «распознаванием» интенции и «размытостью» ответа, вызванной спецификой текста: когда мы имеем дело с интервью, чётко разделить вопросы и ответы не всегда получается, т.к.

интервью – это сложная срежессированная система, в основе которой лежит диалогическое единство (стимул – реакция). В результате ответ из базы знаний может лишь частично совпадать с вопросом пользователя. Тем не менее, мы считаем, что наша модель осуществляет корректную выдачу результатов, при условии схожести пользовательского вопроса с вопросом из базы знаний, и она имеет потенциал использования в электронных региональных СМИ.

При дальнейших разработках следует учесть ряд особенностей и внести следующие изменения:

- добавить систему автопополнения базы знаний вопросов и ответов, если пользователь не удовлетворён выдачей, что приведёт к постоянному автообновлению векторной модели системы;
- необходимо привлечь дополнительные средства семантического анализа: для выявления парадигматических связей в русскоязычных вопросах можно использовать сервис RuWordNet (ruwordnet.ru), а для выявления вопросов-парафразов можно прибегнуть к использованию нейронных сетей. В статье (см. в частности, Митрофанова 2020) авторы как раз уделяют внимание вопросам выявления парафразов в русскоязычных корпусах;
- возможно привлечение тематических моделей типа BTM (biterm topic modelling) для коротких текстов для сравнения тематических лемм пользовательского вопроса и вопроса-эталона;
- в дальнейшем возможно сравнение Word2vec модели с моделью doc2vec, в котором вектор абзаца помогает предсказывать слова из данного вопроса во всех локальных контекстах. Это поможет понять более глобальный контекст.

Список использованной литературы

1. Апресян, Ю.Д. Идеи и методы современной структурной лингвистики. – М. : Просвещение, 1966. – 301 с.
2. Багдасарян, Э.Ю. Интервью как разновидность диалогического общения // Вестник МГОУ. – 2016. – №3 Серия : Лингвистика. – С. 15-21.
3. Белоусов, К.Е. Модельная лингвистика и проблемы моделирования языковой реальности // Вестник ОГУ. – 2010. – № 11 (117). – С. 94-97.
4. Блох, М.Я., Поляков, С. М. Строй диалогической речи / М. Я. Блох, С. М. Поляков. – М. : Прометей, 1992. – С. 7-63.
5. Бокова, А.С. Медиаэффекты жанра интервью в современной этножурналистике // Студенческая наука и XXI век. – 2018. – Т. 15, № 1(16), Ч. 1. – С. 172-173.
6. Бычковская, Н.В. Заголовок массмедийного интервью в аспекте лингвистической прагматики // Liberal Arts in Russia. – 2016. – Vol. 5, No. 1. – С. 58-63.
7. Валгина, Н.С. Активные процессы в современном русском языке: уч. пособие. – М. : Логос, 2003. – 304 с.
8. Верещагин, Е.М., Костомаров, В.Г. Язык и культура : Лингвострановедение в преподавании русского языка как иностранного. – 4-е изд., перераб. и доп. – М. : Рус. яз., 1990. – 246 с.
9. Елецкая, О.В., Тараканова, А.А. Психолого-педагогическая диагностика детей с нарушениями речи : учебно-методич. пособие / О.В. Елецкая, А.А. Таранова. – СПб.: ЛГУ им. А. С. Пушкина, 2012. – 311 с.
10. Есенина, О.А., Щербатых, Е.Ю. Реплики-стимулы в диалогических единствах с оценочным компонентом (на материале современных англоязычных интервью) // Ярославский педагогич. вестн. – 2014. – Т. 1, №2. – С. 134-139.
11. Жбанкова, Е.А. Модели семантического анализа в вопросно-ответной системе [Электронный ресурс]. – URL : https://events.rudn.ru/event/45/papers/306/files/702-ZhbankovaElena.pdf (дата обращения : 14.03.2021).
12. Захаров В.П., Мочалова А.В., Мочалов В.А. Вопросно-ответные системы. Некоторые проблемы автоматической обработки текста. – Петрозаводск : ПИН, 2015. – 40 с.
13. Иванова, И.В. Жанр интервью: формы бытования и языковые особенности : автореф. дисс. … канд. филол. наук. – Астрахань : Астрахан. гос. ун-т. – 2009. – 23 с.
14. Ильченко, С.Н. Интервью в журналистском творчестве : учеб. пособие / С.Н. Ильченко. – СПб. : СПбГУ, 2003. – 93 с.
15. Исакова, Т.Н. Функционирование жанра интервью на современном этапе // Система ценностей современного общества. – 2009. – №5-2. – С. 19-22.

16. Коготкова, С.С. Функции разделительных вопросов в деловом интервью-диалоге // Вестн. РУДН. – 2013. – №2. – С. 101-105.
17. Кожунова, О.С. Моделирование лексической семантики в задачах компьютерной лингвистики // Системы и средства информатики. – 2012. – №1, Т. 22. – С. 86-109.
18. Колесниченко, А.В. Практическая журналистика : учеб. пособие [Электронный ресурс]. – М. : Изд-во Моск. ун-та, 2008. – URL : http://www.evartist.narod.ru/text28/0034.htm (дата обращения : 22.02.2021).
19. Кохан, Н.А. К вопросу о процессе семантической ассимиляции лексики с национально-культурным компонентом // Вестн. Чуваш. ун-та. – 2007. – №4. – С. 181-185.
20. Кошкаров, А.В., Рыбак, К.В. Обзор современного состояния интеллектуальных вопросно-ответных систем // Международный научный журнал «ВЕСТНИК НАУКИ». – 2020. – № 6 (27) Т. 1. – С. 202-205.
21. Кутний, А.И. Интервью в региональных СМИ : типология и специфика жанра // Известия высших учебных заведений. Северо-Кавказский регион. Общественные науки. – 2012. – №6. – С. 130-135.
22. Лободенко, Л.К. Лексико-стилистические особенности информационных жанров медиатекстов интернет-сми региона // Вестник ЮУрГУ. – 2016. –Т. 13, №1. – С. 32-37.
23. Лукина, М. Технология интервью : технология интервью : учеб. пособие для вузов / М. Лукина. – 2-е изд., доп. – М. : Аспект Пресс, 2012. – 192 с.
24. Морозов, И.Ю. Информационное моделирование в лингвистике // Информационные модели в лингвистике : сборник статей / Ред. сост. И.Ю. Морозов. – Омск : Изд-во ОмГПУ, 2001. – С. 36-40.
25. Никифорова, О. О. Речевые особенности жанра интервью (на примере анализа интервью с министром образования Германии Йоганной Ванкой) // Актуальные вопросы филологических наук (II): Матер. Междунар. науч. конф. (г. Чита, июль 2013 г.). – Чита : Молодой ученый, 2013. – 92 с.
26. Панова, Е.Ю. Креативность в профессиональном журналистском образовании : вызовы эпохи vs стандарты // Вестн. Челяб. гос. ун-та. – 2017. – Вып. 109, №11 (407). – С. 43-47.
27. Прибылых, С.Р., Лотова, Н.К. Якутские топонимы лингвистические характеристики и педагогические возможности // Научно-метод. журнал «Концепт». – 2019. – №6. – С. 36-44.
28. Сабитова, А.М., Овчинникова, Н.И. Новостные материалы в региональной печати: жанровые особенности // Международ. научно-исслед. журнал. – 2013. – № 3-2 (10). – С. 22-24.

29. Современная этноязыковая ситуация в Республике Саха (Якутия) : социопсихолингвистический аспект / Васильева Р.И. [и др.] ; отв. ред. П.А. Слепцов. – Новосибирск : Наука, 2013. – 250 с.
30. Соколянская, Н.Н. Проблемы региональной лингвистики (лексикографический аспект). – Магадан : Изд-во Северного междунар. ун-та, 2006. – 85 с.
31. Солганик, Г.Я. Язык современной публицистики / Г.Я. Солганик. – М. : ФЛИНТА, 2017. – 232 с.
32. Соловьёв, А.А., Пескова, О.В. Построение вопросно-ответной системы для русского языка : модуль анализа вопросов // Новые информационные технологии в автоматизированных системах, 2010. – М. : МГТУ им. Н.Э. Баумана. – С. 41-49.
33. Тезаурус русского языка RuWordNet : официальный сайт [Электронный ресурс]. – URL : https://ruwordnet.ru/ru (дата обращения : 27.04.2021).
34. Формальные модели и системы в вычислительной лингвистике / Д. Ш. Сулейманов и [др.] / Под ред. П.И. Соснина, О.А. Невзоровой – Академия наук РТ, Институт прикладной семиотики АН РТ. – Казань : 2016. – 187 с.
35. Черноморова, Т.С., Воробьёв, С.П. Классификация и принципы построения систем вопросно-ответного поиска // Технические науки Technical Sciences, 2020. – №8. Т.6. – С. 145-156.
36. Шатохина, С.И. Проблема первого вопроса в интервью (по материалам журнала «Gala биография» // Современные проблемы науки и образования №2 : материалы конференции. – 2006. – №2. – С. 61-62.
37. Штофф, В.А. Моделирование и философия / В.А. Штофф. – М. ; Л.: Наука, 1966. – 304 с.
38. Щитова Д. А. Интервью как способ создания имиджа // Вестн. Томского гос. ун-та. Филология. – 2012. – Сер. 2. — С. 34.
39. Якутск Вечерний : Vecherniy.com [Электронный ресурс]. – URL : http://www.vecherniy.com (дата обращения : 15.04.2021).
40. ЯСИА – Якутия, все главные новости Якутии и Якутска [Электронный ресурс]. – URL : https://ysia.ru (дата обращения : 05.01.2021).
41. Abraham Ittycheriah. A Statistical Approach For Open Domain Question Answering // Advances in Open Domain Question Answering. Springer Netherlands, 2006. – Part 1. V ol.32. – pages 35-69.
42. Burger, J. and others. Issues, tasks and program structures to roadmap research in question & answering (Q&A) // NIST DUC Vision and Roadmap Documents, 2001. – URL : http://www.nlpir.nist.gov/projects/duc/roadmapping.html (дата обращения : 14.03.2021).

43. DeepPavlov [Электронный ресурс] : официальный сайт DeepPavlov – URL : https://deeppavlov.ai (дата обращения : 03.02.2021).
44. Exactus Интеллектуальный метапоиск в Интернете. ИСА РАН. [Электронный ресурс]: официальный сайт Exactus. – URl: http://exactus.ru/ (дата обращения : 13.05.2020).
45. Kemachart Kemavuthanon, Osamu Uchida. Integrated Question-Answering Systems for Natural Disaster Domains Based on Social Media Messages Posted at the Time of Disaster // Knowledge Management, Trust and Communication in the Era of Social Media, 2020. pages 1-14. – URL : https://www.mdpi.com/2078-2489/11/9/456/htm (дата обращения : 22.03.2021).
46. Paulo Cavalin and others. Building a Question-Answering Corpus using Social Media and News Articles // International Conference on Computational Processing of the Portuguese Language, 2016. pages 1-7. – URL : https://www.researchgate.net/publication/304189920_Building_a_Question-Answering_Corpus_Using_Social_Media_and_News_Articles (дата обращения : 20.03.2021).
47. Shivani G. Aithal, Abishek B. Rao, Sanjay Singh. Automatic question-answer pairs generation and question similarity mechanism in question answering system // Applied Intelligence, 2021. pages 1-15. – URL : https://www.researchgate.net/publication/350707864_Automatic_question-answer_pairs_generation_and_question_similarity_mechanism_in_question_answering_system (дата обращения : 22.03.2021).
48. START [Электронный ресурс] : официальный сайт START. – URL : http://start.csail.mit.edu/index.php (дата обращения : 11.05.2020).
49. Vadim Gudkov, Olga Mitrofanova, Elizaveta Filippskikh. 2020. Automatically Ranked Russian Paraphrase Corpus for Text Generation. pages 54-59. – URL : https://www.aclweb.org/anthology/2020.ngt-1.6.pdf (дата обращения : 09.05.2021).
50. Ykt.ru – главный якутский портал. Новости Якутска и Якутии, форумы, блоги, знакомства, погода, объявления, афиша [Электронный ресурс]. – URL : https://www.ykt.ru (дата обращения : 13.04.2021).

Словари и справочники

51. Активный словарь русского языка / В.Ю. Апресян [и др.]. – М. : Языки славянской культуры, 2014. – Т. 1. – 408 с.
52. Большой энциклопедический словарь (БЭС) [Электронный ресурс] / А.М. Прохоров. – 2-е изд., перераб. и доп. – М. : Большая Российская Энцикл., 1997. – URL https://dic.academic.ru/dic.nsf/enc3p/200678 (дата обращения : 10.05.2021).

53. Брызгунова, Е.А. Русская грамматика : [В 2 т.] / Н.Ю. Шведова [и др.]. – Т. 2 : Синтаксис [Электронный ресурс] / Е.А. Брызгунова [и др.]. – М. : АН СССР, Ин-т рус. яз., 1980. – URL : http://rusgram.narod.ru/2591-2640.html#2629 (дата обращения : 27.01.2020).
54. Лингвистический энциклопедический словарь (ЛЭС) [Электронный ресурс] / Г.В. Якушева, В.Н. Ярцева. – М. : Большая Рос. энцикл., 2002. – 707 с. – URL : http://lingvisticheskiy-slovar.ru (дата обращения : 06.05.2020).
55. Ляшевская, О.Н., Шаров, С.А. Частотный словарь современного русского языка (на материалах Национального корпуса русского языка). – М. : Азбуковник, 2009. – 1087 с.
56. Розенталь, Д.Э., Теленкова, М.А. Словарь-справочник лингвистических терминов [Электронный ресурс]. – 2-е. изд. – М. : Просвещение, 1976. – URL : https://dic.academic.ru/dic.nsf/lingvistic/751 (дата обращения : 06.03.2021).
57. Русский ассоциативный словарь : в 2 Т. [Электронный ресурс] / Ю.Н. Караулов [и др.]. – М., 2002. – URL : http://tesaurus.ru/dict/ (дата обращения: 15.04.2021).

Приложение А Листинг программы вопросно-ответной системы для регионального издания ЯСИА

```
pip install pymorphy2  # устанавливаем библиотеку для морфологического анализа на этот сервер

from openpyxl import load_workbook  # библиотека для открытия Excel файлов
import gensim  # библиотека для построения w2v моделей
from gensim.models import Word2Vec
import pymorphy2
import re  # регулярные выражения
import numpy as np  # преобразование слов в векторы
from numpy import dot  # следующие два подмодуля нужны для подсчёта косинусной близости
from numpy.linalg import norm

wb = load_workbook('Q&A.xlsx')  # открываем файл

sheet = wb['Sheet1']  # открываем нужный лист

questions = []    # будем сюда добавлять вопросы и ответы
answers = []

for i in range(2, 1305):    # считываем информацию из ячеек и добавляем в списки
  a_value = 'A'+str(i)  # пишем имя ячейки (A2, A3 и т.д.)
  b_value = 'B'+str(i)
  q = sheet[a_value].value  # обращаемся к ячейке по имени и достаём значение
  a = sheet[b_value].value
  questions.append(q)  # присоединяем полученные вопросы и ответы к спискам, которые мы создали чуть выше
  answers.append(a)

swl = open('swl.txt', encoding='utf8')    # стоп-слова
swlist = []
for i in swl:
    i = i.replace('\n','')
    swlist.append(i)
swl.close()

final_pr = [] # пустой список для токенизированных и лемматизированных вопросов
morph = pymorphy2.MorphAnalyzer() # создаём функцию для морфологического анализа
for i in questions: # одновременно обрабатываем по одному вопросу
  lst = re.findall(r'[0-9A-Яёа-яёһөүҥ\-]+', i) # в интервью есть и несколько вопросов и ответов на якутском
  new_list = []  # сюда будем добавлять леммы слов из вопросов
  for w in lst: # нормальная форма для токенов
    k = morph.parse(w)[0] # из словаря OpenCorpora берём морф. хар-ки (первого слова по частоте — нужное нам)
    k = k.normal_form
    if k not in swlist:
      new_list.append(k)
  final_pr.append(new_list)

model = Word2Vec(final_pr, size=300, window=4, min_count=1, workers=4)  # создание w2v модели
model.init_sims(replace=True)  # дополнительная нормализация векторов
model.save('qa_system_w2v.model')
```

```
vector_questions= []  # сюда будем добавлять векторные представления вопросов
for quest in final_pr:
  vectors_mean = []  # сюда добавляем матричное представление векторов w2v в NumPy
  for y in quest:  # для каждого слова в "мешке слов" из одного вопроса
    try:
      vector = model.wv[y]  # перевод слова в вектор
      vectors_mean.append(vector)  # присоединение к общему вектору
    except:
      pass
  try:
    average = sum(vectors_mean)/len(vectors_mean)  # средний вектор нашего вопроса
    vector_questions.append(average)  # присоединяем векторную форму к общему списку для сравнения пользовательского вопроса
    #с базой данных
  except:
    average = np.zeros_like(vector_questions[-1])  # если не удалось перевести вопрос в векторную форму,
      # то берём последний созданный вектор и создаём вектор из нулей того же размера, что и последний вектор
    vector_questions.append(average)

def qasystem():  # функция ввода вопроса и сравнения
  user = input('Введите Ваш вопрос, посвящённый Якутии: ')
  tokens = re.findall(r'[0-9а-яёһҕөүҥґ\-]+', user)
  new_list = []
  for token in tokens:
    k = morph.parse(token)[0]
    k = k.normal_form
    if k not in swlist:
      new_list.append(k)
  user_arr = []
  for y in new_list:
    try:
      vector = model.wv[y]
      user_arr.append(vector)
    except:
      pass
  try:
    average_user = sum(user_arr)/len(user_arr)
  except:
    average_user = np.zeros_like(vector_questions[-1])
  all_cosines = []  # каждый вопрос из базы данных будем сравнивать с пользовательским и считать сходство по косинусной мере
  for i in vector_questions:
    cos_sim = dot(i, average_user)/(norm(i)*norm(average_user))
    all_cosines.append(cos_sim)
  max_index = all_cosines.index(max(all_cosines))  # индекс максимального значения из списка all_cosines
  # print(max_index)
  final_a = 'Ответ из базы данных: '+answers[max_index]  # полученный индекс применяем к списку с ответами и выводим сам ответ
  return final_a
```