На правах рукописи

КОРЯКОВЦЕВ Михаил Андреевич

МОДЕЛЬ «ПРЕДСКАЗАНИЯ» СЛОВОФОРМЫ НЕФОРМАЛИЗОВАННОЙ ЧАСТИ ТЕКСТА ЭЛЕКТРОННОГО РЕЗЮМЕ

Специальность 10.02.21 – Прикладная и математическая лингвистика

А В Т О Р Е Ф Е Р А Т
диссертации на соискание ученой степени кандидата филологических наук

Тюмень
2017

Работа выполнена на кафедре английской филологии и перевода Института филологии и журналистики федерального государственного автономного образовательного учреждения высшего образования «Тюменский государственный университет»

Научный руководитель:	**Табанакова Вера Дмитриевна** доктор филологических наук, профессор кафедры английской филологии и перевода ФГАОУ ВО «Тюменский государственный университет»
Официальные оппоненты:	**Шереметьева Светлана Олеговна** доктор филологических наук, профессор кафедры лингвистики и перевода ФГАОУ ВО «Южно-Уральский государственный университет (национальный исследовательский университет)»
	Влавацкая Марина Витальевна доктор филологических наук, профессор кафедры иностранных языков ФГБОУ ВО «Новосибирский государственный технический университет»
Ведущая организация:	**ФГБОУ ВО «Пермский государственный национальный исследовательский университет»**

Защита состоится 22 ноября 2017 года в 9:00 на заседании диссертационного совета Д 212.274.15 по защите диссертаций на соискание учёной степени кандидата филологических наук при Тюменском государственном университете по адресу: 625003, г. Тюмень, ул. Володарского, 6, корпус 1, ауд. 211.

С диссертацией можно ознакомиться в библиотеке ИБЦ ФГАОУ ВО «Тюменский государственный университет» по адресу: 625003, г. Тюмень, ул. Семакова, 18, а также на официальном сайте ТюмГУ, код доступа: http://d21227415.utmn.ru/defenses

Автореферат разослан «___» _____ 2017 г.

Ученый секретарь
диссертационного совета Д 212.274.15,
доцент, кандидат филологических наук *Д. В. Шапочкин*

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Реферируемая диссертационная работа посвящена формализации текста электронного резюме и разработке модели предсказания словоформы для обеспечения оптимизации заполнения электронного резюме на сайтах трудоустройства. Разработка и апробация модели предсказания словоформы была проведена с использованием текстового материала рубрики «Обязанности, функции, достижения» электронных резюме сайта трудоустройства HeadHunter. Данное исследование выполнено в русле современной прикладной лингвистики на стыке нескольких областей знаний: теории коммуникации, теории текста, теории формальных грамматик, комбинаторной лингвистики и автоматической обработки текста.

Актуальность данного исследования и **степень разработанности** обусловлены следующими положениями:

1. Сайты трудоустройства, на которых люди оставляют свои резюме, а организации – вакансии, становятся все более популярными. Их количество для русскоязычного сегмента сети Интернет по данным сервиса «Яндекс.Каталог» составляет порядка 200 на середину 2015 г, а количество электронных резюме, размещенных на некоторых сайтах, превышает миллион. При заполнении резюме от соискателя требуется вручную заполнить электронные бланки на каждом из интересующих его сайтов. При этом каждый сайт трудоустройства предлагает свою структуру и средства для заполнения этих бланков. Тем не менее, большая часть текстового содержимого электронного резюме остается неформализованной. Это требует от соискателя сочинения и ручного ввода некоторого текста по заданной им же самим структуре.

2. В настоящее время имеющиеся программы по автоматическому заполнению электронных бланков и сами сайты трудоустройства не предоставляют соискателю вспомогательные средства для заполнения всех частей электронного резюме. Программы по заполнению электронных бланков, например RoboForm, пытаются автоматически заполнить электронные бланки на указанных сайтах трудоустройства, но в то же время требуют написания первичного резюме в текстовом файле, либо в них самих.

3. Последние работы по изучению резюме выполнены в трёх направлениях. В работе С. А. Ярцева [Ярцев 2012] резюме рассматривается как форма деловой коммуникации соискателя и работодателя при устройстве на работу. В работе О. В. Тойкиной [Тойкина 2014] резюме исследуется с точки зрения его жанровых особенностей в рамках документального

типа текста; в докладах Т. В. Качаевой и В. С. Южикова [Качаева Южиков 2007] и А. В. Сафронова [Сафронов 2008] электронное резюме является объектом задачи классификации в процессе автоматизированной обработки текста. Все перечисленные направления занимаются исследованием характеристик уже готового текста резюме.

Объектом исследования является текстовое содержимое электронного резюме на популярных сайтах трудоустройства, а **предметом исследования –** параметры модели предсказания словоформы неформализованного текстового содержимого электронного резюме.

Гипотеза исследования заключается в том, что оптимизация написания неформализованной части текста электронного резюме может быть осуществлена при помощи предсказания словоформы в предложении с помощью синтаксически связанных пар словоформ, выбранных из конечного, заранее сформированного набора таким образом, что одна из словоформ в паре присутствует в уже введённой соискателем части предложения.

Материалом диссертации послужили:

– 228 рубрик, образующих 14 электронных бланков трёх наиболее популярных в русскоязычной части сети Интернет сайтов трудоустройства: «HeadHunter», «SuperJob» и «Работа.Ru».
– 4495 текстов рубрики «Функции, обязанности, достижения» электронного резюме сайта «HeadHunter» в сфере деятельности «Программирование, Разработка» общим объемом 836005 словоупотреблений.

Целью исследования является разработка модели предсказания словоформы неформализованной части текстового содержимого электронного резюме.

Выдвинутая гипотеза исследования и поставленная цель решаются выполнением следующего ряда **задач**:

1) Ознакомиться со средой функционирования текста электронного резюме и описать коммуникативную ситуацию его публикации и поиска на сайтах трудоустройства.

2) Исследовать электронное резюме как особый вид текста.

3) Описать структуру текста электронного резюме и установить формализованность его элементов.

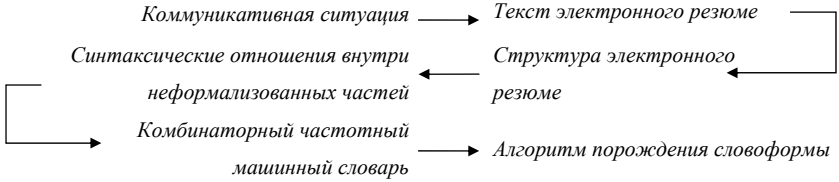
4) Провести синтаксический анализ неформализованной части электронного резюме с целью установления бинарных отношений между словоформами.

5) Построить комбинаторный частотный машинный словарь. Данный словарь должен включать частоты словоформ, частоты синтаксических пар и расположение словоформ в этих парах.

6) Разработать модель предсказания словоформы неформализованной рубрики с использованием комбинаторного частотного машинного словаря.

7) Разработать программный комплекс оптимизации заполнения электронного резюме.

Таким образом, ход исследования может быть представлен следующей схемой:



Теоретическую основу и методологическую базу составили:

- работы в области теории коммуникаций К. Черри (1972), Р. О. Якобсона (1985), В. В. Красных (2001), А. В. Соколова (2002), М. М. Назарова (2002), М. А. Василика (2003), А. П. Панфиловой (2004), В. А. Масловой (2008), Ф. И. Шаркова (2010);
- работы по теории общего и специального текста Н. И. Жинкина (1982), А. И. Новикова (1983), А. С. Герда (1996), Е. С. Кубряковой (2001), Н. С. Валгиной (2003), В. Д. Табанаковой (2004, 2009), И. Р. Гальперина (2007), М. Н. Кожиной (2008), В. П. Москвина (2012);
- работы в области теории формальных грамматик Н. Хомского (1958, 2005), И. А. Мельчука и А. В. Гладких (1969, 1970);
- работы по фреймовому анализу Э. Гоффмана (1974), Ч. Филлмора (1976, 1981, 1982, 1985, 1988), М. Минского (1979), M. Petruck (1986, 1996), А. Н. Баранова (1987, 2001), Т. А. ванн Дейка (1989), Н. Н. Болдырева (2000), Л. А. Нефедовой (2003), О. В. Соколовой (2007), Ж. В. Никонова (2008), С. Л. Мишлановой (2010, 2012), Ю. С. Верхотуровой (2012), В. С. Кавицкой (2013);
- работы в области комбинаторной лингвистики и лексикографии В. В. Морковкина (1970, 1977), Ю. Н. Караулова (1976), М. В. Влавацкой (2007, 2009, 2011);
- работы в области автоматизации обработки текста Дж. Хопкрофта (2002), И. Сегаловича (2003);
- работы в области word prediction (предсказание словоформы) Briandias (1959), Lesher (1999), Lassila (1989, 1998, 2001), MacKenzie (1995, 1998, 2001, 2002), Trnka (2008).

Для решения поставленных задач использовался комплексный **метод исследования**, включающий в себя *общенаучные методы*: описание, анализ, сравнение, классификацию, а также *частнонаучные методы*: дефиниционный анализ, лингвистическое моделирование, методику фреймового моделирования, синтаксический анализ по методу грамматики зависимостей (dependency grammar), методы построения и анализа алгоритмов.

Проведенное исследование позволяет вынести на защиту следующие **положения**:

1. Взаимодействие работодателя и соискателя через посредника – информационно-поисковую систему сайта трудоустройства – описывается функциональной моделью электронно-поисковой коммуникации. В электронно-поисковой коммуникации «работодатель – сайт трудоустройства – соискатель» было выделено восемь коммуникативных процессов, каждый из которых описывается циркулярной моделью коммуникации. В качестве сообщений выступают поисковые запросы, электронные бланки, электронные резюме и электронные вакансии.

2. Содержимое электронного резюме рассматривается как особый вид текста. Текст электронного резюме обладает формальной иерархической структурой. Эта структура представлена в виде фрейма, узлами и слотами которого являются рубрики. Рубрики-слоты подразделяются на формализованные и неформализованные. Дифференцирующим параметром является синтаксическая оформленность: рубрика-слот, содержащая номинацию, является формализованной, а рубрика-слот, содержащая высказывание, – неформализованной.

3. Формализация текста рубрики «Обязанности, функции, достижения» на синтаксическом уровне предполагает построение дерева зависимостей. Каждое ребро в дереве зависимостей отражает упорядоченное бинарное отношение между словоформами, которое мы называем синтаксической парой. Синтаксическая пара словоформ и её частотные характеристики составляют структуру словарной статьи комбинаторного частотного машинного словаря.

4. Алгоритм предсказания словоформы неформализованного текстового содержимого электронного резюме опирается на последовательное использование частотных параметров синтаксических пар комбинаторного частотного машинного словаря. Алгоритм реализован в программном комплексе оптимизации заполнения электронных резюме на сайте трудоустройства HeadHunter. Эффективность предложенного алгоритма устанавливается относительно простого набора текста и двух алгоритмов

предсказания словоформы: автодополнение слова по частотному словарю и предиктивный ввод на основе префиксного дерева (trie).

Научная новизна диссертационной работы заключается в том, что:

1) Впервые предметом исследования становятся параметры модели предсказания словоформы текста электронного резюме.

2) Впервые в процессе моделирования содержательной структуры электронных резюме анализируются их текстовые признаки.

3) Исследование структуры электронных резюме происходит с использованием первичного, по отношению к текстовому содержимому электронного резюме, объекта – электронного бланка.

4) Впервые при описании смысловых единиц в тексте резюме применяется параметр формализованности.

5) Впервые для предсказания словоформы используется специализированный комбинаторный частотный машинный словарь синтаксических пар.

Теоретическая значимость Теоретическая значимость диссертационного исследования определяется тем, что оно вносит вклад в развитие таких областей прикладной лингвистики, как теория электронной коммуникации, теория специального текста в электронном формате, комбинаторная лексикография и автоматическая обработка текста. В частности выстраивается модель коммуникативного процесса на сайтах трудоустройства, устанавливается степень формализованности текста электронного резюме, доказывается положение о формализации текстового содержимого электронного резюме с помощью комбинаторного словаря, обосновываются принципы алгоритма предсказания словоформы текстового содержимого электронного резюме.

Практическая ценность диссертационного исследования определяется применением фреймового моделирования для описания структуры текстового содержимого электронного резюме, составлением комбинаторного частотного машинного словаря, обеспечивающего оптимизацию ввода текста электронного резюме, разработкой программного комплекса оптимизации заполнения электронного резюме.

Апробация работы. Основные положения диссертационного исследования отражены в 6 публикациях общим объемом 3,69 п. л. (из которых 3,69 печатных листов выполнены единолично автором), в том числе в 3 статьях, вышедших в свет в изданиях, рекомендуемых ВАК при Министерстве образования и науки РФ. Результаты диссертационного исследования освещались автором на конференциях: «Прикладная лингвистика в науке и образовании ALPAC report – полвека после разгрома» (Санкт-

Петербург, 24-26 ноября 2016), «Экология языка на перекрёстке наук» (Тюмень, 20-21 ноября 2014), «Фундаментальная наука и технологии – перспективные разработки» (North Charleston, USA, 29-30 сентября 2014).

Диссертационная работа состоит из введения, трёх глав, заключения, списка литературы и трёх приложений. Объём диссертации составляет 145 страниц. В диссертации содержится 42 рисунка и 6 таблиц. Список литературы состоит из 156 источников (из них 30 источников – на иностранных языках).

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во **введении** указывается цель работы, обосновывается её актуальность, теоретическая и практическая значимость, описываются задачи исследования, объект, предмет и материал исследования, используемые методы исследования, указываются положения, выносимые на защиту, и приводятся сведения об апробации результатов работы.

В первой главе «Текст электронного резюме в коммуникативной ситуации «работодатель – сайт трудоустройства – соискатель» электронное резюме было исследовано как особый вида текста в рамках электронно-поисковой коммуникации. В ходе исследования было определено понятие электронно-поисковой коммуникации, построена модель коммуникативной ситуации «работодатель – сайт трудоустройства – соискатель» и описаны признаки текста электронного резюме.

В **разделе 1.1** в работах Ф. И. Шаркова, П. А. Панфиловой, М. А. Василика, А. В. Соколова и К. Черри были рассмотрены три вида коммуникации – *социальная*, *массовая* и *электронная (компьютерная)* – и выявлены их существенные и отличительные признаки.

По результатам анализа и обобщения трёх видов коммуникаций было сформулировано определение понятия электронно-поисковой коммуникации: общение, целенаправленная передача информации от человека к человеку (или группы людей к другой группе людей) с помощью языковых средств с использованием информационно-поисковой системы в качестве средства связи в процессе их познавательно-трудовой деятельности. У электронно-поисковой коммуникации были выделены четыре отличительных признака:

1) наличие трёх участников коммуникации: отправитель, получатель и посредник

2) отправитель и получатель могут представлять собой как некоторую группу людей, так и отдельных индивидов, при этом получатель представляет собой безадресную аудиторию

3) в качестве посредника и средства связи в электронно-поисковой коммуникации выступает информационно-поисковая система

4) общение отправителя с получателем происходит в неинтерактивном режиме, в то время как отправитель с посредником и получатель с посредником общаются в интерактивном режиме

В **разделе 1.2** коммуникативная ситуация взаимодействия работодателя и соискателя была описана в виде составной функциональной модели коммуникативного процесса «работодатель – сайт трудоустройства – соискатель». Данная модель представляет собой совокупность 8 интерактивных коммуникативных процессов, описывающих взаимодействие между сайтом трудоустройства и соискателем или работодателем: 1) получение электронного бланка вакансии, 2) публикация вакансии, 3) поиск электронной вакансии, 4) получение электронной вакансии, 5) получение электронного бланка резюме, 6) публикация электронного резюме, 7) поиск электронного резюме, 8) получение электронного резюме. Схема коммуникативной ситуации представлена на рисунке 1.

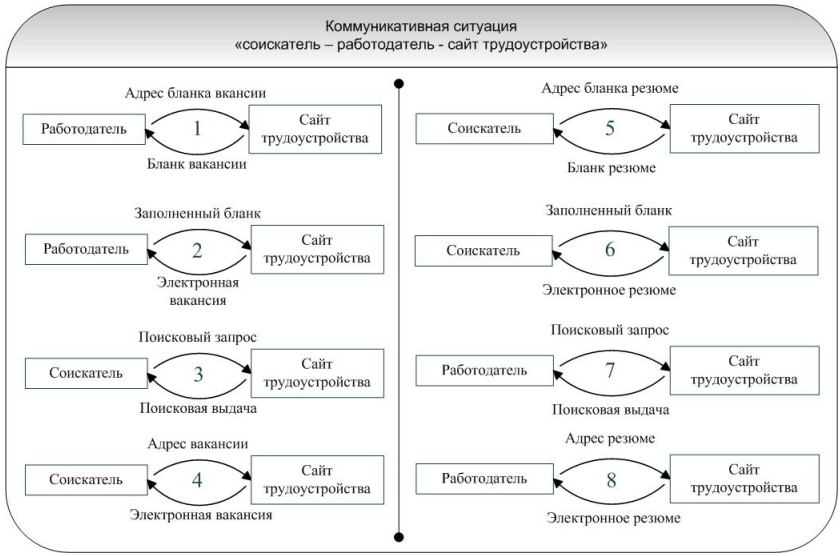


Рис. 1. Схема коммуникативной ситуации
«соискатель – работодатель – сайт трудоустройства»

Каждый из восьми коммуникативных процессов описан циклической моделью и включает в себя следующие ключевые структурные элементы: коммуникаторы, кодирование и декодирование, канал и сообщение. Канал представляет собой сеть Интернет, кодирование и декодирование сообщений – процесс визуализации HTML тегов средствами веб-браузеров. Сообщения включают в себя *адрес страницы, поисковый запрос, бланки резюме и вакансий, резюме, вакансии* и *поисковые выдачи*. В качестве коммуникаторов выступают работодатель, соискатель и сайт трудоустройства. При этом: р*аботодатель* – это организация, заинтересованная в найме сотрудника, отвечающего определённым критериям, на конкретную должность. С*оискатель* – это индивид, имеющий определённые знания, обладающий определёнными навыками и желающий устроиться работать на определённую должность. С*айт трудоустройства* – это информационно-поисковая система, доступная в сети Интернет. Сайт трудоустройства предлагает следующие возможности: размещение объявлений о найме и поиске работы и поиск среди объявлений о найме и поиске работы.

В разделе 1.3 были рассмотрены признаки текста электронного резюме. Эти признаки были получены в результате анализа определений понятия *текст* И. Р. Гальперина, Е. С. Кубряковой и В. П. Москвина. Было установлено, что тексту электронного резюме присущи следующие 12 признаков:

1. Письменная форма. Электронное резюме предстает как последовательность символов на родном (русском) языке.

2. Наличие заголовка. Электронное резюме имеет два типа заголовков: общий заголовок резюме и подзаголовок (заголовок логического блока).

3 и 4. Завершённость и **прагматическая направленность** электронного резюме диктуются основной функцией резюме – самопрезентацией соискателя. Завершённым считается такое резюме, в котором соискатель указал все важные составляющие резюме.

5. Стилистическая согласованность обусловлена наличием следующих важных составляющих резюме: данные об образовании, сведения о приобретённых навыках и умениях, опыте работы.

6. **Информативность** представлена наличием содержательно-фактуальной информации (пол, дата рождения, заработная плата и т.д.) и содержательно-концептуальной информации (автор представляет себя как опытного и достойного кандидата на заявленную должность).

7. Членимость. Электронному резюме присуща только объемно-прагматическая членимость. Единицей членение был выбран логический

блок – тематически организованная часть текста (образование, уровень, место обучения и т. д.).

8. Связность электронных резюме проявляется наличием трёх средств когезии: традиционно-грамматических, логических и стилистических. *Традиционно-грамматические* средства когезии представлены повсеместным употреблением персонального дейксиса в имплицитной форме. *Логические* и *стилистические* средства когезии присутствуют внутри отдельных логических блоков электронных резюме и представляют собой списки и перечисления.

9. Самодостаточность. Самодостаточным является содержание отдельных логических блоков электронных резюме.

10. Интертекстуальность, т. е. совпадения электронных резюме по форме и содержанию, проявляется в как рамках одного сайта трудоустройства, так и на разных сайтах.

11. Протяжённость. Тексты резюме различаются по своему объёму – от 60 до 1600 словоупотреблений. При этом, даже тексты с минимальным объёмом являются завершёнными и стилистически согласованными, т. е. в них присутствуют все необходимые логические блоки.

12. Адресат. Адресатом в электронных резюме является работодатель.

На основании 10 из 12 признаков текста были заданы три параметра формальной модели электронного резюме: ***замкнутость, обобщённость и преемственность***. *Замкнутости* формальной модели электронного резюме означает, что она не требует для своего функционирования других моделей. *Обобщенность* даёт возможность построения общей формальной модели электронных резюме нескольких сайтов трудоустройства. *Преемственность* указывает, что формальная модель должна содержать модели электронного резюме, заложенные в электронных бланках создателями сайтов трудоустройства. Связь признаков текста и параметров формальной модели представлена в таблице 1.

Таблица 1

Связь признаков текста и параметров формальной модели

Замкнутость	Обобщённость	Преемственность
самодостаточность	письменная форма	наличие заголовка
интертекстуальность	прагматическая направленность	завершённость
адресат	информативность	членимость
	членимость	связность
	интертекстуальность	

Во второй главе диссертации «Структурно-функциональная модель электронного резюме» описывается построение модели электронного резюме на двух уровнях: фреймовом и синтаксическом.

Вначале был проведён обзор работ по фреймовому моделированию в когнитивной лингвистике (фреймовая семантика) и компьютерной лингвистике (базы знаний). Когнитивный подход представлен в работах зарубежных (M. Minsky, C. Fillmore, M. Petruck), так и отечественных (Н. Н. Болдырев, В. А. Маслова, В. З. Демьянков, А. Н. Баранов) авторов. По результатам обзора были установлены следующие различия в способах представления фрейма в когнитивной и компьютерной лингвистике:

1) В когнитивной лингвистике объект описывается единственным фреймом, имеющим разветвлённую структуру и разнообразие связей. В компьютерной лингвистике каждая составляющая или компонент объекта описан с помощью отдельного фрейма, образующих таксономию.

2) Когнитивная лингвистика предполагает наличие значения и явное описание его ограничений только в терминальных узлах фрейма. Связи между узлами, при этом, рассматриваются как неявные ограничения. В компьютерной лингвистике каждый фрейм содержит в себе слоты, которые включают в себя набор фасетов.

В диссертационной работе для описания электронного резюме был использован когнитивный фреймовый подход, предложенный M. Petruck. Структурными элементами фрейма электронного резюме послужили узлы, связи между узлами, слоты и ограничения слотов.

В качестве **узлов** фрейма электронного резюме используются вербализованные части электронных бланков. Элементы управления считаются **слотами**, а выбранный или введенный в них текст – наполнением слота. Узлы и слоты, состоящие из вербализованных частей и элементов управления, в работе называют ***рубриками.*** Рубрики в электронном резюме имеют между собой только один тип **связи** – подрубрика.

В качестве **ограничений** в слотах была использована характеристика *план выражения*. Характеристика *план выражения* может принимать три значения: *выбор, словосочетание и текст*. Значение характеристики определяется по методу ввода у соответствующего слоту элемента управления. Всего было выделено два метода ввода: *выбор* и *ввод текста*. Метод ввода *выбор* означает, что с помощью элемента управления

соискатель выбирает слова или словосочетания из предлагаемого списка заданных значений. Метод ввода *ввод текста* означает, что соискатель вводит некоторый текст с помощью клавиатуры. Методу ввода *выбор* соответствует значение плана выражения *выбор*, тогда как методу ввода *ввод текста – словосочетание* и *текст*. Значение плана выражения для метода ввода *ввод текста* зависит от того, какую синтаксическую оформленность принимает содержимое элемента управления. В случае если его содержимое является номинацией, то план выражения будет принимать значение *словосочетание*, если же его содержимое – высказывание, то план выражения – *текст*.

С помощью описанного подхода были получены фреймы трёх сайтов трудоустройства: HeadHunter, SuperJob и Работа.Ru. Отбор перечисленных сайтов трудоустройства был произведен с использованием тематического индекса цитирования в качестве показателя популярности. Для иллюстрации в тексте автореферата на рисунке 2 представлена часть фрейма электронного резюме HeadHunter. Фреймы электронного резюме HeadHunter, SuperJob и Работа.Ru полностью представлены в приложениях А, Б и В диссертации.

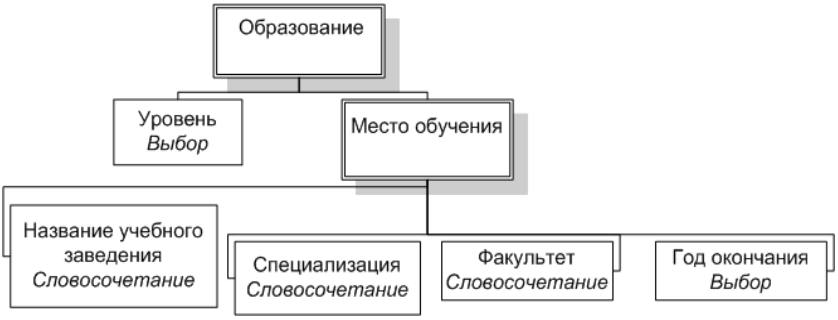


Рис. 2. Узлы и слоты фрейма, соответствующие электронному бланку сайта трудоустройства HeadHunter

Рубрики фреймов трёх сайтов трудоустройства были подвергнуты количественному и качественному анализу с целью выявления наибольшей по объёму и, в то же время, универсальной для всех сайтов рубрики.

Количественный анализ показал, что рубрика *Обязанности, функции, достижения* является самой большой по количеству словоупотреблений.

Её объём в среднем составляет 47% всех словоупотреблений. Качественный анализ рубрик *Обязанности, функции, достижения* и *Обязанности и достижения* трёх сайтов трудоустройства показал, что они носят универсальный характер. Это связано с тем, что данные рубрики на всех трёх сайтах трудоустройства имеют одинаковый лексический контекст. Следовательно, формализацию достаточно проводить на материале одного сайта трудоустройства.

Синтаксический анализ текста рубрики *Обязанности, функции, достижения* проводился на материале электронных резюме сайта трудоустройства HeadHunter. Он включил в себя два предварительных этапа – извлечение текста и разбиение текста рубрики на простые части – и один заключительный – систематизация синтаксических отношений.

Синтаксический анализ состоял в построении неразмеченного дерева зависимостей для каждого предложения. В качестве синтаксических отношений в дереве зависимостей использовались поверхностно-синтаксические отношения, представленные в работе И. А. Мельчука.

Систематизация синтаксических отношений была выполнена в виде комбинаторного частотного машинного словаря. Комбинаторный частотный машинный словарь Cd представляет собой множество пар словоформ и их параметров. Словоформы в паре связаны синтаксическим отношением, являющимся ребром деревьев зависимостей. Синтаксическая пара словоформ выполняет одновременно несколько функций. Она, с одной стороны, образует обязательную для словарей комбинаторного типа сочетаемостную зону, а, с другой стороны, является заголовочной единицей.

Микроструктура комбинаторного частотного словаря была задана в виде набора компонентов:

$$Cd \in Cd = < f_l, f_r, fq_a, fq_l, fq_r, d, ai_l, ai_r >$$

где

f_l и f_r – пара синтаксически связанных словоформ;

fq_a – **абсолютная частота употребления синтаксической пары** в используемом языковом материале. Служит для выбора синтаксической пары в качестве единицы предсказания среди других синтаксических пар, имеющих общую словоформу;

fq_l, fq_r – **абсолютная частота употребления левой и правой словоформы** синтаксической пары. Служит для предсказания словоформы

в отсутствии контекста, но при наличии введённой части порождаемой словоформы;

d **– доминантность левой словоформы синтаксической пары.** Она указывает, что левая словоформа может быть использована для предсказания первой словоформы в предложении.

ai_l, ai_r – **индекс направления** ai_f словоформы f служит для обозначения направления предполагаемого расширения текста для этой словоформы. Значение этого индекса лежит в диапазоне [–1; 1]. Отрицательные значения индекса направления словоформы указывают, что большая часть синтаксически связанных словоформ располагается справа от текущей словоформы. Положительные значения, наоборот, указывают, что большая часть синтаксически связанных словоформ располагается слева от текущей словоформы. Значения –1 и 1 указывают, что синтаксически связанные словоформы располагаются только справа или слева соответственно.

Индекс направления ai_f может быть получен следующим образом:

$$ai_f = \frac{rfreq_f - lfreq_f}{rfreq_f + lfreq_f}$$

где

$rfreq_f$ – абсолютная частота словоформы f в качестве правой части синтаксической пары;

$lfreq_f$ – абсолютная частота словоформы f в качестве левой части синтаксической пары.

Часть таблицы, соответствующей комбинаторному частотному машинному словарю в базе знаний модели предсказания словоформы представлена в таблице 2.

Таблица 2

Часть таблицы комбинаторного частотного машинного словаря

f_l	f_r	d	fq_a	ai_l	ai_r	fq_l	fq_r
1	*2*	*3*	*4*	*5*	*6*	*7*	*8*
разработка	приложений	Да	157	-0,985	1	1288	194
программного	обеспечения	Нет	148	-1	0,276	175	155
разработка	сайтов	Да	142	-0,985	1	1288	299
разработка	системы	Да	102	-0,985	0,147	1288	236

Окончание табл. 2

1	*2*	*3*	*4*	*5*	*6*	*7*	*8*
разработка	обеспечения	Да	101	-0,985	0,276	1288	155
разработка	систем	Да	85	-0,985	1	1288	234
разработка	проектов	Да	79	-0,985	1	1288	163

В третьей главе «Модель «предсказания» словоформы текстового содержимого рубрики «Обязанности, функции, достижения» описаны существующие модели предсказания словоформы, предлагается алгоритм предсказания словоформы с использованием частотного машинного словаря и сравниваются существующие и предложенный алгоритмы.

В **разделе 3.1** даётся обзор решений проблемы оптимизации ввода на мобильных устройствах. В рамках этой оптимизации рассматриваются модели word frequency lists (частотные списки словоформ), Trie (префиксное дерево) и FussyTree и сравниваются лингвистические модели, лежащие в их основе. В модели на основе частотных списков словоформ используются словоформы и их частоты в некотором корпусе текстов. Лингвистическая модель префиксного дерева *Trie* представляет собой набор N-грамм символов и их частот, тогда как в модели *FussyTree*, являющейся наследником *Trie*, осуществляется переход с уровня N-грамм символов к N-граммам словоформ.

В **разделе 3.2** представлен алгоритм порождения словоформы на основе комбинаторного частотного машинного словаря.

Проблема, которую решает предложенный алгоритм, сформулирована следующим образом:

Пусть некоторый текстовый документ может быть представлен в виде последовательности словоформ $w_1, w_2, w_3 \ldots w_N$. *Необходимо для имеющейся последовательности словоформ* $w_1, w_2, w_3, \ldots w \ldots w_{N-1}, w_N$ *предсказать такую словоформу* w, *что вероятность её корректности будет максимальной.*

Алгоритм предсказания словоформы включает две операции предсказания *Suggest* и $Suggest_0$ в зависимости от текущего слова w (частично введенная словоформа) или w_0 (отсутствие ввода). Операция *Suggest* рассматривается как предсказание словоформы в тексте методом восстановления текущей словоформы, а $Suggest_0$ – как предсказание новой словоформы.

Операции *Suggest* и $Suggest_0$ в своей работе используют следующую общую для них обоих логику построения списка предполагаемых словоформ:

1) В каждый отдельный момент времени суждение о выборе отдельной предполагаемой словоформы из пары словоформ в словарной статье может быть сделано при помощи не более чем одной существующей в предложении словоформе.

2) В первую очередь должны быть обработаны словоформы, находящиеся слева от точки предсказания, а затем справа от неё.

3) В последнюю очередь должны быть представлены словоформы вне зависимости от словоформ левого или правого окружения.

Словоформы, входящие в полученный в результате работы этих алгоритмов список, отсортированы четырьмя способами:

1) Сортировка по типу окружения, для которого предполагается словоформа: использовано левое окружение, использовано правое окружение и без использования окружения.

2) Сортировка по удалённости синтаксической пары от порождаемой словоформы: от ближайших словоформ к наиболее удалённым.

3) Сортировка по частоте употребления в массиве текстов, на основе которого сформирован комбинаторный частотный машинный словарь.

4) Сортировка по месту следующего предсказания словоформы: слева или справа от текущей словоформы.

Эффективность работы алгоритма, т. е. то, в какой позиции в списке находится нужная соискателю словоформа, представленных алгоритмов предсказания полностью зависит от выборки, на основе которой формируется комбинаторный частотный машинный словарь.

В **разделе 3.3** было проведено сравнение трёх алгоритмов предсказания словоформы: на основе частотного списка, на основе комбинаторного частотного машинного словаря и FussyTree. Параметром сравнения был выбран Key Savings (KS), определённый следующим образом:

$$KS = \frac{keystrokes_{normal} - keystrokes_{with\ prediction}}{keystrokes_{normal}} * 100\%$$

где

$keystrokes_{normal}$ – количество нажатий для ввода некоторого текста;

$keystrokes_{with\ prediction}$ – количество нажатий для ввода того же самого текста с помощью системы предсказания слов.

Сравнение включило в себя три этапа ввода текста:

1) Ввод одного случайного предложения из обучающей выборки (выборке текстов, используемой при формировании словаря).

2) Ввод случайного предложения, не входящего в обучающую выборку. При этом электронное резюме, в которое входит этот текст, подано на должность в той же сфере деятельности, что и электронные резюме в обучающей выборке.

3) Ввод предложения, не входящего в обучающую выборку и поданного на должность в другой сфере деятельности.

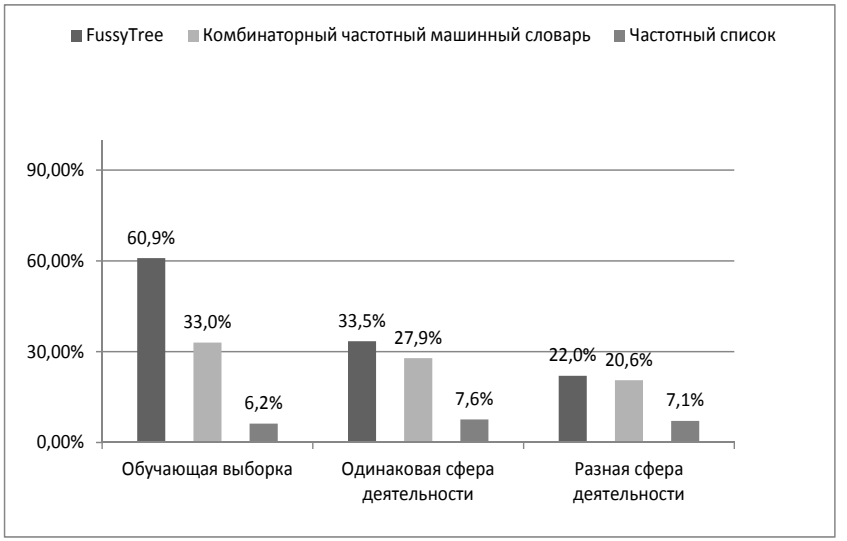


Рис. 3. Показатели KS для алгоритмов предсказания словоформы

По результатам оценки алгоритмов предсказания словоформы по показателю KS, представленному в диаграмме на рисунке 3, в диссертационной работе был сделан следующий вывод. Предложенный алгоритм предсказания словоформы на основе комбинаторного частотного машинного словаря намного превосходит алгоритм на основе частотных списков словоформ. По сравнению с алгоритмом на основе префиксного дерева он, с одной стороны, в значительной степени отстаёт от него на обучающей выборке, а, с другой стороны, сравним с ним при использовании во время написания новых текстов как в той же

сфере деятельности, так и в отличной от сферы деятельности обучающей выборки.

Таким образом, выдвинутая гипотеза была подтверждена: предсказание словоформы неформализованной части текста электронного резюме с использованием синтаксических пар, полученных из существующих электронных резюме, является возможным.

В **Заключении** приводятся общие итоги и выводы по проведённому исследованию.

В **приложениях** представлены три фреймовые модели текста электронного резюме соответствующие электронным резюме сайтов трудоустройства HeadHunter, SuperJob и Работа.Ru. Так же приложение включает в себя 90 предложений послужившие материалом апробации алгоритма предсказания словоформы. Из них 30 предложений входят в обучающую выборку, 30 предложений из электронных резюме в сфере деятельности «Программирование, разработка» и 30 предложений в сфере деятельности «Юриспруденция».

Основные положения диссертации отражены в следующих публикациях.

В изданиях, рекомендованных ВАК РФ:

Коряковцев, М. А. Определение степени формализованности текста резюме на сайтах трудоустройства [Текст] / М. А. Коряковцев // Вестник Челябинского государственного педагогического университета. – 2015. – № 9. – С. 146-150

Коряковцев, М. А. Текстовые признаки электронных резюме [Текст] / М. А. Коряковцев // Вестник Челябинского государственного педагогического университета. – 2016. – № 5. – С. 167-175

Коряковцев, М. А. Использование конечных автоматов для извлечения текстовых данных электронных резюме сайтов трудоустройства HEADHUNTER, SUPERJOB и РАБОТА.RU [Текст] / М. А. Коряковцев // Высшее образование сегодня. – 2016. – № 11. – С. 18-22

Прочие публикации по теме диссертации:

Коряковцев, М. А. Применение современных компьютерных технологий при формировании минимальной выборки текстов резюме интернет-ресурса Superjob.ru [Текст] / М. А. Коряковцев // Фундаментальная наука и технологии – перспективные разработки Материалы

IV международной научно-практической конференции. – North Charleston (USA): CreateSpace, 2014. – С. 181.

Коряковцев, М. А. Электронный бланк резюме как прототип фрейм-мовой модели «электронное резюме» [Текст] / М. А. Коряковцев // Наука и образование в социокультурном пространстве современного общества. Сборник научных трудов по материалам Международной научно-практической конференции: в 3-х частях. – Смоленск: ООО «НОВАЛЕНСО», 2016. – С. 130-134

Коряковцев, М. А. Особенности построения онтологии электронного резюме при разработке кадровой информационной системы [Текст] / М. А. Коряковцев // Труды VIII Международной научной конференции «Прикладная лингвистика в науке и образовании. ALPAC REPORT – полвека после разгрома». 24-26 ноября 2016. г., Санкт-Петербург. – СПб.: ООО «Книжный Дом», 2016. – С. 61-66

Подписано в печать 19.09.2017. Тираж 120 экз.
Объем 1,0 уч. изд. л. Формат 60х84/16. Заказ 602.

Издательство Тюменского государственного университета
625000, г. Тюмень, ул. Семакова, 10
Тел./факс (3452) 59-74-68, 59-74-81
E-mail: izdatelstvo@utmn.ru