На правах рукописи

ТОЛПЕГИН Павел Владимирович

АВТОМАТИЧЕСКОЕ РАЗРЕШЕНИЕ КОРЕФЕРЕНЦИИ МЕСТОИМЕНИЙ ТРЕТЬЕГО ЛИЦА РУССКОЯЗЫЧНЫХ ТЕКСТОВ

Специальность 05.13.17 – Теоретические основы информатики

АВТОРЕФЕРАТ

диссертации на соискание ученой степени кандидата технических наук

Москва – 2008

Работа выполнена в Вычислительном центре им. А.А. Дородницына Российской академии наук, Отдел математических проблем распознавания и методов комбинаторного анализа

Научный руководитель: доктор физико-математических наук, профессор
Рязанов Владимир Васильевич

Официальные оппоненты:

доктор технических наук
Зеленков Юрий Григорьевич

доктор технических наук, профессор
Местецкий Леонид Моисеевич

Ведущая организация: **Институт системного анализа Российской академии наук** (ИСА РАН)

Защита диссертации состоится 18 декабря 2008 г. в 14 час. на заседании диссертационного совета Д 002.017.02 Вычислительного центра им. А.А. Дородницына Российской академии наук по адресу: 119333, Москва, ул. Вавилова, д. 40.

С диссертацией можно ознакомиться в библиотеке ВЦ РАН.

Автореферат разослан 17 ноября 2008 г.

Ученый секретарь
диссертационного совета,
д.ф.-м.н., проф. В.В. Рязанов

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы

Всемерное распространение и совершенствование информационных технологий вызвали мощный импульс к исследованиям в области анализа текстовых данных. При извлечении информации из текста на естественном языке (ЕЯ) важным условием качества понимания является отождествление повторно упоминаемых объектов. Актуальной задачей представляется разработка специализированных моделей распознавания и алгоритмических средств по переводу линейной структуры текста в структуру, отражающую сложные смысловые отношения между объектами Мира.

В представленной работе исследуется одна из центральных проблем автоматической обработки текстов (АОТ) – проблема автоматического разрешения анафорических связей. Предлагаются алгоритмы установления кореферентных связей, приводятся практические результаты для информационно-новостных текстов. Функциональная сторона разрешения анафоры*, как этапа ЕЯ-анализа, заключается в установлении зависимостей между объектами (именными и другими группами), расположенными в простых предложениях (клаузах) на протяжении единицы текста. Указанная проблема исследовалась на больших корпусах ЕЯ-текстов с использованием методов математической теории распознавания.

Работа с корпусами текстов представляется актуальной по ряду причин. Во-первых, в 60 – 90-е гг. XX в. различные виды знаний закладывались в ЭВМ вручную в форме частных правил, при этом не использовались средства их автоматического извлечения из корпусов текста. Во-вторых, достоверные числовые характеристики и показатели от работы с корпусом можно получить, оперируя с большими объёмами текста.

Особый интерес автоматическое разрешение анафорических связей (в частности – кореференции местоимений) представляет при проектировании систем автоматического машинного перевода, информационного поиска и разработке вопросно-ответных систем. Последние могут быть также полезны для расширения смыслового представления текста. Вместе с тем, несмотря на востребованность практических систем автоматического определения кореферентных связей, развитых разработок для русского языка в настоящее время не существует. На этом фоне, однако, продолжают совершенствоваться зарубежные разработки текстового анализа.

Учитывая изложенное, компьютерная обработка русскоязычного текста, осуществляющая автоматическое определение кореферентных связей между анафо́ром (далее в нашем случае – местоимением третьего лица) и стоящим ранее по тексту неким объектом Мира (антецеде́нтом), представляется актуальной задачей.

Теоретической и методологической основой исследования послужили труды отечественных и зарубежных ученых в

---

* использование выразительных свойств языка, которые могут быть корректно проинтерпретированы только в контексте (с учётом предшествующего фрагмента текста)

области математической и прикладной лингвистики, машинного перевода Н.Д. Арутюновой, Т.В. Булыгиной, Дж. Гандел, А.А. Кибрика, Л.Н. Иорданской, Дж. Николс, Е.В. Падучевой, Е.В. Рахилиной, А.С. Чехова, А.Д. Шмелева, М.И. Откупщиковой, Р.В. Миткова, В.Г. Гака, И.А. Муравьевой, О.Ю. Богуславской, Ю.С. Мартемьянова, А.В. Гулыги, Е.М. Вольф, З.М. Шаляпиной, И.И. Ревзина, работы в области машинного обучения и распознавания образов Ю.И. Журавлева, В.Л. Матросова, К.В. Рудакова, В.В. Рязанова, О.В. Сенько, исследования специалистов в области искусственного интеллекта и автоматической обработки текстов Д.А. Поспелова, Г.С. Осипова, В.Ф. Хорошевского, Ю.Г. Зеленкова, А.Н. Аверкина, А.И. Эрлиха и др.

Цель и задачи исследования

Цель исследования – разработка подхода автоматического определения кореферентных связей для русского языка, основанного на анализе корпусов текстов с использованием методов теории распознавания.

Для достижения поставленной цели были решены следующие задачи исследования:

– систематизация формальных средств выражений анафорических связей и зависимостей для местоимений третьего лица;

– выделение и формализация признаков, влияющих на референциальный выбор;

– разработка методов и алгоритмов формирования новых признаков по неразмеченным корпусам текстов без привлечения средств семантики, логики и знаний о Мире;

– разработка алгоритмов и программ, основанных на подходах и методах теории распознавания, обеспечивающих автоматическое определение кореферентных связей между анафо́ром и антецедентом.

– создание экспериментальной программной среды для аккумуляции статистической информации о референциальном выборе реального антецедента для местоимения третьего лица;

– создание корпуса русскоязычных ЕЯ-текстов, размеченных экспертом на предмет кореферентных связей между анафо́ром и антецедентом, а также размеченных автоматически морфологическими, синтаксическими и первично-семантическими анализаторами;

– разрешение задачи установления кореферентных связей для информационно-новостных текстов, оценка влияния каждого из факторов на корректность определения кореферентных связей при принятии решения о референциальном выборе и поиск минимальных признаковых подпространств.

Объект исследования – сфера автоматического определения кореферентных связей, а также факторы и признаки, влияющие на этот процесс, их анализ при помощи методов машинного обучения.

Предмет исследования – методы и алгоритмы, формирующие признаковое пространство при определении кореферентных связей; свойства

признаков, генерируемых упомянутыми методами и алгоритмами; модели распознавания кореферентных связей.

Материалами исследования послужили тексты электронных новостных изданий. Общий объём автоматически проанализированных текстов составил более 140 Мбайт.

Научная новизна. Современная деловая проза (в т.ч. информационно-новостные и др. тексты) русского языка не изучались до настоящего времени на предмет выявления закономерностей в референциальном выборе местоимений третьего лица при помощи методов машинного обучения и распознавания образов. В работе впервые применены подходы к изучению закономерностей кореферентных связей с применением методов машинного обучения и распознавания образов. Разработаны и апробированы новые методы и алгоритмы, «компенсирующие» нехватку семантических знаний, знаний «о Мире» и логических правил из неразмеченных корпусов текстов, новые алгоритмы синтеза корпусных признаков, а также предложены новые модели распознавания кореферентных связей.

Теоретическая значимость исследования заключается в разработке подхода для автоматического разрешения анафорических связей, создании методов синтеза корпусных признаков и моделей распознавания кореференции.

Практическая значимость состоит в использовании разработанных алгоритмов определения кореферентных связей при разрешении ана́форы в задачах машинного перевода, автоматического реферирования текстов, извлечения информации в поисковых и диалоговых системах и других автоматических системах искусственного интеллекта в части АОТ.

На защиту выносятся следующие положения:

1. методы анализа неразмеченных корпусных ресурсов (источников большого объёма ЕЯ-текстов) и результаты их применения в задаче разрешения кореференции местоимений;
2. алгоритмы по расширению признакового пространства в задаче разрешения кореференции русскоязычных текстов:
   – алгоритм вычисления оценок степени встречаемости одушевлённости для валентностей русскоязычных глаголов;
   – алгоритм синтаксической деривации;
   – алгоритм определения конфликтующих антецедентов;
   – алгоритмы по формированию корпусных оценок степени встречаемости гипотетического антецедента и глагольной группы, управляющей анафо́ром;
3. алгоритм некореферентности анафо́ра с гипотетическим антецедентом и алгоритм некореферентности местоимений;
4. результаты анализа признаков при установлении кореферентных связей и минимальные подпространства признаков;

5. модели распознавания для разрешения ана́форы местоимений третьего лица в русскоязычных текстах;

6. реализация моделей распознавания в виде программной среды, обеспечивающей дружественный интерфейс для работы эксперта по разметке текстов на предмет анафорических связей и автоматическую аккумуляцию признаков;

7. модель подготовки и обработки размеченных ЕЯ-текстов с целью выявления закономерностей и значимых систем признаков;

8. результаты испытания модели на размеченном корпусе информационно-новостных текстов (свыше 2000 фрагментов текстов объёмом, превышающим 3 Мбайт).

Апробация. Основные научные выводы и результаты исследования докладывались и обсуждались на:

(1) международной конференции «Диалог 2006» – Компьютерная лингвистика и интеллектуальные технологии (Бекасово, 31 мая – 4 июня 2006 г.);

(2) 10-ой национальной конференции по искусственному интеллекту с международным участием КИИ-06 (Обнинск, 25-28 сентября 2006 г.);

(3) научно-технической конференции «Информационные технологии в бизнесе» (Москва, ГУ ВШЭ, 2006);

(4) международной конференции «Диалог 2007» – Компьютерная лингвистика и интеллектуальные технологии (Бекасово, 30 мая – 3 июня 2007 г.);

(5) 7-ой международной конференции «Информационное общество, интеллектуальная обработка информации, информационные технологии», 24-26 октября 2007 г. НТИ-2007. (Москва, ВИНИТИ РАН).

Публикации. По теме диссертации опубликовано 18 работ, общим объёмом 197 стр. Из них 2 – в издании из списка, рекомендуемых ВАК Минобрнауки России – журнал «Информационные технологии» (№№ 8,9, 2006 г.).

Составляющие диссертационной работы поддержаны:

(1) конкурсом ведущих научных школ «НШ-5833.2006.1» 2006 г.: «Развитие фундаментальных математических основ и алгоритмического аппарата для решения сложных задач интеллектуального анализа данных, распознавания и прогнозирования» (исполнитель проекта);

(2) научной стипендией ООО «Яндекс» 2004-2005 гг.: «Разработка, создание и внедрение процедуры апостериорной оценки качества поиска на основе поведения пользователей» (рук. проекта);

(3) грантом РФФИ № 06-06-80464-а 2006 г.: «Разработка и реализация методов семантического и прагматического анализов ЕЯ-текстов русского языка» (рук. проекта);

(4) научной стипендией ООО «Яндекс» 2006-2007 гг.: «Формирование нечётких мер для валентностей русскоязычных глаголов» (рук. проекта).

Структура работы. Диссертация состоит из введения, четырёх глав и заключения. Основной текст изложен на 179 стр. при общем объёме 241 стр., включая 3 приложения и библиографию из 181 наименования.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении обоснована актуальность темы диссертации, сформулированы её цели и задачи, описаны методы исследования.

В первой главе («Задача референциального анализа и методы решения») приводится обзор современного состояния в данной области, описание существующих подходов и методов решения задачи разрешения ана́форы, обоснование целесообразности разработки методов её решения на базе теории распознавания с использованием корпусных средств.

Вводится понятие первично-семантического графа, который строится автоматически для каждого отдельного предложения на этапе первичного семантического анализа [Сокирко, 2005]. Первично-семантический граф G – это ориентированное дерево, вершиной которого является глагольная группа (*root*), узлами которого являются члены предложения, а ребрами – валентные связи. Рассматривались следующие основные свойства узлов графа G: $w_i.morpho$ – морфологические характеристики (кроме одушевлённости); $w_i.inf$ – начальная форма; $w_i.anim$ – одушевлённость; $w_i.top$ – управляющий узел. Понятие первично-семантического графа прдеставляется базовым во многих задачах АОТ.

Рассматривается задача референциального анализа – построение когнитивной карты диску́рса (абзаца, смысловой единицы текста). Разрешение анафорических связей (связи местоимения с расположенными до него по тексту объектами Мира) считается главным этапом при построении когнитивной карты и при семантическом связывании серии предложений.

Вводится перечень видов знаний, применяемых для разрешения местоимённой ана́форы (морфологические, синтаксические, семантические, знания диску́рса, знания о Мире).

В [Aone и Bennett, 1995, 1996] предложена идея разрешения ана́форы с использованием алгоритмов распознавания образов – MLR (Machine learning resolver). Задача выявления кореферентной связи между заданным анафо́ром *Анф* и некоторым $ГА_j$ при известном списке конкурирующих гипотетических антецедентов $ГА_1$, $ГА_2$, … , $ГА_j$ решалась в два этапа. Сначала происходило распознавание наличия кореферентной связи между анафо́ром и каждым отдельным $ГА_i$, i=1,2, … ,l, с помощью решающего дерева (алгоритм C4.5 [Quinlan, 1993]) по признаковым описаниям соответствующих пар $<Анф, ГА_i>$. Далее применялись эвристические логические правила для выбора реального антецедента. В случае если пара $<Анф, ГА_i>$ является кореферентной, решающее дерево возвращает тип анафорической связи. Обучение осуществлялось только для тех местоимений, которые были идентифицированы программой автоматически. Корпус для обучения содержал 1971 анафо́р, 1359 из которых были идентифицированы

программой. Точность работы системы (доля правильно определённых кореферентных связей) составила от 83,49 до 88,55%.

Система RESOLVE [McCarthy и Lehnert, 1995] также использовала алгоритм C4.5 – решающие деревья. Вектор признаков, применяемый при обучении, формировался в отдельности для кореферентных и некореферентных пар <*Анф*, $ГА_i$>. Выборка состояла из 322 объектов первого класса (кореферентные пары) и 908 объектов второго класса (некореферентные пары). Признаки содержали информацию о референции к имени собственному, о дублирующих референциях к одной и той же именной группе, метрические признаки (осуществляется ли кореференция к ГА, находящемуся в том же предложении, что и рассматриваемый Анф). Корпусные признаки и признаки одушевлённости не использовались. Точность работы системы составила 85,8%.

Подходы [Soon, Ng, Lim, 1999] и [Soon и др., 2001], основанные на алгоритмах C4.5 и C5, соответственно, дополнительно использовали определение семантических классов слов и их групп на основе словаря WordNet. Вектор признаков строился для пар <*Анф*, $ГА_i$> и включал метрические показатели, признак согласованности в роде и числе, семантический класс, показатель имени собственного. Точность составила 68%.

В рамках настоящей работы были проведены эксперименты по разрешению ана́форы для русскоязычных текстов, когда для описания объектов распознавания (соответствующих пар <*Анф*, $ГА_i$>) использовались 14 базовых признаков: метрика, морфология, упрощённый синтаксис, упрощённая семантика и др. Точность распознавания с применением различных подходов на тестовой выборке составила около 62%. Анализ ошибочных контекстов показал, что величина ошибки в обучении преимущественно зависит от выразительных средств языка, которые не подчиняются правилам, отражённым в данной системе признаков. Делается вывод о необходимости расширять систему признаков, привлекая новые источники статистической информации (корпусные данные), и совершенствовать модели распознавания с целью повышения точности разрешения ана́форы.

Во второй главе («Вычисление признаков в задаче установления кореферентных связей») приводится описание методов генерации новых признаков с применением корпусно-ориентированных средств и алгоритмов поиска оптимальных признаковых подпространств. Это позволяет расширить исходное признаковое пространство за счёт анализа корпусов текстов и частично компенсировать нехватку данных, относящихся к компетенции логики, семантики, знаний о Мире. Отсутствие комплексных семантических и онтологических ресурсов продиктовало необходимость в использовании корпуса текста большого объёма, размеченного синтаксическими связями, в качестве источника дополнительной информации. В основу подхода по формированию корпусных признаков положена идея оценки степени встречаемости глагольной группы и гипотетического антецедента в

синтаксически размеченном корпусе [Ido Dagan и Alon Itai, 1990, 1991]. Предлагается вычислять численные оценки степени встречаемости глагольной группы, управляющей анафо́ром, попарно с каждым из гипотетических антецедентов, используя неразмеченные корпуса текстов большого объёма, а также оценки степени одушевлённости валентностей русскоязычных глаголов (корпусные признаки), признак конфликтности антецедентов.

Иллюстрация корпусного признака:

В автомобиль, Иван, встроил блокиратор, коробки переключения передач. Теперь ***его*** *сложно угнать.*

По данным поиска шаблона «$ГА_i$»+«*ГГ*» в имеющемся неразмеченном корпусе текстов «*угнать автомобиль*» встречается в 124 раза чаще, чем «*угнать блокиратор*». А «*угнать Ивана*» – в 43 раза реже, чем «*угнать автомобиль*». Соответствующие частоты используются для вычисления серии корпусных признаков.

Вычисление признаков в задаче установления кореферентных связей с использованием корпусных средств

(1) Корпусные признаки основаны на оценке степени встречаемости глагольной группы (*ГГ*), управляющей анафо́ром, попарно с каждым из гипотетических антецедентов ($ГА_i$) в неразмеченном корпусе текстов. Были предложены следующие корпусные признаки, реализующие идею сочетаемости *ГГ* и $ГА_i$:

корпусный признак №1. Степень сочетаемости *ГГ* и $ГА_i$ с предлогом на расстоянии до *t* слов, в прямом или обратном порядке, без ограничений на морфологию;

корпусный признак №2. Степень сочетаемости *ГГ* и $ГА_i$ с предлогом контактно, в прямом или обратном порядке, $ГА_i$ находится в падеже анафо́ра;

корпусный признак №3. Степень сочетаемости *ГГ* и $ГА_i$ контактно в прямом или обратном порядке с уточняющим (присвязочным) словом.

(2) Признаки, характеризующие «степень одушевлённости» валентностей русскоязычных глаголов. Существуют контексты, референциальный выбор в которых требует наличия информации о степени одушевлённости у исследуемого глагола в заданной валентности.

Пример 1. Контекст, требующий знаний одушевлённости

Маша купила машину. ***Она её*** *любит.*

Предложен метод по автоматическому извлечению знаний об одушевлённости из корпусов неразмеченных текстов в целях формирования признаков степени одушевлённости.

Метод предполагает первичную обработку входных текстов процессором «Диалинг» с расширенным семантическим интерфейсом.

Например, для предложения «*Фёдор вышел на террасу*» анализатор возвращает следующую структуру:

№ предложения; № узла; w_i; $w_i.inf$;$w_i.morpho$+$w_i.anim$; № управляющего узла; val_k

```
1;0; ФЕДОР;ФЕДОР;С,имя,од,мр,им,ед,;1;SUB;
1;1; ВЫШЕЛ;ВЫЙТИ;Г,дст,нп,св,прш,мр,ед,;ROOT;;
1;2;НА ТЕРРАСУ;ТЕРРАСА;С,но,жр,вн,ед,;1;TRG-PNT;
```

Для каждого идентифицированного глагола v_j ($v_j.morpho$=*ГЛАГОЛ*) устанавливаются подчинённые ему существительные $\{w_i: \exists val_k(v_j, w_i)\}$ путём обхода первично-семантического графа как ориентированного дерева. Для каждого w_i определяются морфологические характеристики, в частности – морфологическая одушевлённость. Таким образом, раздельный подсчёт числа одушевлённых (*од*) и неодушевлённых (*но*) w_i для рассматриваемого входного условия $\{v_j, val_k\}$ или $\{v_j, val_k, w_i.morpho\}$ позволяет сформировать оценки степени одушевлённости и алгоритмы их вычисления.

Оценка степени встречаемости неодушевлённости для глагола в рамках заданной валентности:

$$\mu_1(v_j, val_k) = \frac{\sum_i \left| w_i.anim = "но" \right|}{\sum_i \left| w_i.anim = "од" \right| + \sum_i \left| w_i.anim = "но" \right|}$$, где суммирование по *i*: $val(v_j, w_i) = val_k$, $w_i.top = v_j$.

Оценка степени встречаемости неодушевлённости для глагола в рамках заданных валентности и морфологических характеристик:

$$\mu_2(v_i, val_k, w_j.morpho) = \frac{\sum_i \left| w_i.anim = "но" \right|}{\sum_i \left| w_i.anim = "од" \right| + \sum_i \left| w_i.anim = "но" \right|}$$, где суммирование по *i*: $val(v_j, w_i) = val_k$, $w_i.top = v_j$, $(w_i.morpho = v_j.morpho)$.

Оценка степени встречаемости, показывающая насколько типично существование подчинённых прецедентов в той или иной валентности с определёнными морфологическими характеристиками у заданного глагола по сравнению с другими морфологическими характеристиками того же глагола и той же валентности:

$$\mu_3(v_i, val_k, v_j.morpho) = \frac{\sum_i \left| w_i.morpho = v_j.morpho \right|}{\sum_i \left| w_i.morpho \right|}$$, где суммирование по *i*: $val(v_j, w_i) = val_k$, $w_i.top = v_j$.

Значения оценок степени одушевлённости $\mu_1(v_i, val_k)$, $\mu_2(v_i, val_k, w_j.morpho)$ и $\mu_3(v_i, val_k, w_j.morpho)$ используются для расширенного признакового пространства в качестве числовых признаков с индексами BJ61, BK62, BL63. На базе указанных числовых признаков формировались бинарные признаки (BY76, BZ77, CA78): значение признака устанавливалось равным «1» в случае достижения признаком максимального значения среди

антецедентов для рассматриваемого анафо́ра, и равным «0» – в противном случае.

Для ранее приведённого контекста, местоимения *она* и *её* находятся в разных валентных зависимостях от глагола *любить*: SUB и CONTEN соответственно. Оценка степени одушевлённости μ_1 для этого глагола и валентности SUB составляет 70%, а для валентности CONTEN – 39%. Данный признак является важным для установления кореферентности следующих пар: *она* и *Маша*, *её* и *машина*.

В процессе практической реализации метода «начитано» 80 Мбайт ЕЯ-текстов, по которым автоматически сформирован словарь, содержащий глаголы с оценкой одушевлённости.

(3) Разработан алгоритм определения конфликтующих антецедентов, позволяющий формировать отдельный признак в объективно неоднозначных контекстах.

Пример 2. Омонимия выбора из конфликтующих антецедентов

*Сложно понять логику$_?$ организации$_?$ речи$_?$, **которая**$_i$ нарушена у многих пациентов с поражением левого полушария.*

Три антецедента: *логика*, *организация* и *речь* считаются конфликтующими.

Узлы w_1 и w_2 семантического графа G являются конфликтующими и значение признака CJ87 для каждого узла устанавливается равным «1», если выполнены следующие условия:

1. $w_1.morpho=w_2.morpho$ – у рассматриваемых узлов совпадают род и число, и, таким образом, они могут одновременно выступать в качестве кандидатов для кореференции;
2. $w_1.morpho=w_2.morpho$=СУЩ;
3. $\exists\, val_k(w_1, w_2)$ – между узлами существует связь.

При невыполнении хотя бы одного из условий значений признака CJ87 устанавливается равным «0».

Определение числа гипотетических антецедентов

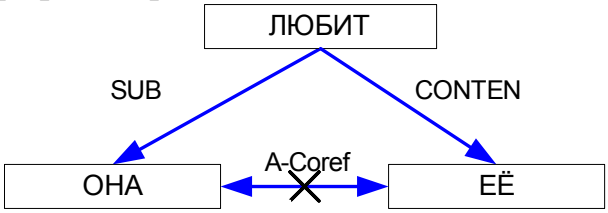
При определении числа антецедентов, являющихся кандидатами при разрешении кореференции для заданного анафо́ра, применялись следующие критерии их отбора.

(1) Два узла w_2 и w_3 ориентированного дерева G являются соподчинёнными, если $\exists\, val_x(w_1, w_2)$ и $\exists\, val_y(w_1, w_3)$. Для участия в референциальном выборе допускаются существительные, совпадающее в роде и числе с анафо́ром, не соподчинённые рассматриваемому анафо́ру, а также причастия или прилагательные, совпадающее в роде и числе с анафо́ром, не имеющие зависимых существительных (напр., *отдыхающие*).

Областью поиска кандидатов для кореференции являются q предложений, стоящих ранее по тексту и предложение, в котором расположен анафо́р. Считается, что для каждого местоимения существует антецедент и он единственный.

(2) Применяется ограничительный признак наличия некореферентной связи анафо́ра с гипотетическим антецедентом. Известно, что гипотетический антецедент и анафорическое выражение (кроме возвратного местоимения) не могут быть кореферентны, если они являются соподчинёнными.

(3) Признак наличия некореферентности местоимений: два местоимения не могут быть кореферентны, если они являются соподчинёнными. Граф для предложения «*Она её любит*»:



Нахождение оптимального признакового пространства

В настоящей работе задача установления кореференции сводится к стандартным задачам распознавания. Пусть задано множество $M=\{S\}$ объектов S. Известно, что M является объединением непересекающихся подмножеств $K_i, i=\overline{1,l}$, называемых классами: $M=\bigcup_{i=1}^{l} K_i$, $K_i \bigcap_{i\neq j} K_j = \varnothing$. Дана начальная информация I_0 о разбиении на классы в виде обучающей выборки $S_{m_{i-1}+1},\dots,S_{m_i} \in K_i, i=\overline{1,l}$, m_0=0, m_l=m. Считаем, что заданы признаковые описания $I(S_i)=(x_1(S_i),x_2(S_i),\dots,x_n(S_i))$, $I(S)$ объектов S_i, S с помощью набора n числовых признаков, характеризующих различные свойства объектов. Требуется ответить на вопрос: $S \in K_j ?, j=\overline{1,l}$.

Совокупность векторов-строк описаний объектов из обучающего множества M_0 может быть записана в виде таблицы T_{mnl}, называемой стандартной таблицей обучения, где m – число объектов обучающего множества, n – размерность признакового пространства, l – число классов.

Важной компонентой обучения является нахождение минимальных признаковых подпространств, сохраняющих достигнутую точность распознавания для исходного признакового пространства. В работе были исследованы различные подходы для нахождения указанных подпространств для задач с двумя классами.

(а) *поиск признакового подпространства на базе метода достоверных статистических разбиений*

Производится оценка индивидуальной способности каждого из признаков по разделению объектов двух классов выборки S^1_{ini}, заданной в виде таблицы T_{MNL}. Оценка производилась с использованием *метода оптимальных статистически достоверных разбиений*

[Журавлев, Рязанов, Сенько, 2006]. Для каждого признака X находится такая пороговая точка, которая наилучшим образом разделяет объекты классов K_1 и K_2 на выборке S^1_{ini} . Для оценки степени разделения используется функционал $F = [(\nu_1 - \nu_1^1)^2 m_1 + (\nu_1 - \nu_1^2)^2 m_2] / \nu_1(1-\nu_1)$, где

ν_1 – доля объектов класса K_1 в S^1_{ini};

ν_1^1 – доля объектов класса K_1 в подмножестве S^1_{ini} с $X<\delta$;

ν_1^2 – доля объектов класса K_1 в подмножестве S^1_{ini} с $X>\delta$;

m_1 – число объектов S^1_{ini} с $X<\delta$;

m_2 – число объектов S^1_{ini} с $X>\delta$.

Для каждого признака с помощью перестановочного теста и метода оптимальных разбиений оценивается статистическая значимость различий в распределениях объектов классов K_1 и K_2. Пусть требуется оценить статистическую значимость некоторой закономерности с пороговым значением δ^{ini}_{opt} и оптимальным (максимальным) значением F^{ini}_{opt} функционала F. Генерируется множество из $N0$ случайных выборок $\{S^r_1,...,S^r_{N0}\}$, совпадающих по числу объектов с S^1_{ini}. В каждой из выборок S^r_i осуществляется случайная перестановка меток классов. Для полученной выборки S'^r_i ищется оптимальное пороговое значение δ_{opt} вместе с оптимальным значением F_{opt} функционала F. Данные вычисления повторяются $N0$ раз.

В качестве меры статистической значимости закономерности (p–значения) принимается доля выборок из $\{S^r_1,...,S^r_{N0}\}$, для которых $F_{opt} > F^{ini}_{opt}$.

(б) *выделение значимых признаков на основе тупиковых тестов*

Использовались результаты работы стохастического варианта метода «Голосование по тупиковым тестам». В качестве входных данных используются выходные параметры тупиковых тестов случайных подтаблиц таблицы T_{mnl}: веса и номера признаков. Пусть проведена серия из $N1$ расчетов на обучающей выборке S^1_{ini}.

Определение степени значимости признаков X_i и их ранжирование проводится по формуле:

$$W(X_i) = \frac{\sum_{k=1}^{N} w_k{}^1(X_i)}{N} + \sum_{j=1}^{Y} \frac{w_j{}^2(X_i)}{W}, \text{ где}$$

N – общее число вхождений признака в тупиковые тесты единичной длины;

W – число признаков, участвующих в данном тупиковом тесте для тестов неединичной длины;

Y – общее число вхождений признака в тупиковые тесты неединичной длины;

$w_k{}^1(X_i)$, $w_j{}^2(X_i)$ – веса тупиковых тестов.

(в) *оптимизация признакового пространства на основе логических корреляций*

При минимизации признакового пространства используются логические закономерности классов, найденные по обучающей выборке. Пусть $N(i,j)$ – число одновременных вхождений признаков X_i, X_j в одну закономерность по множеству логических закономерностей P, найденных по данным обучения. Величина $LogCorr(i,j) = 1 - \frac{N(i,j)}{\min(N(i),N(j))}$ называется логической корреляцией признаков X_i и X_j.

Рассматривается задача нахождения кластеров признаков, для которых входящие в них признаки обладают близкими корреляционными свойствами. В качестве меры корреляционной близости рассматривается критерий, основанный на полуметрике:

$$r(i,j) = \sum_{l=1,l\neq i,j}^{N} \left| LogCorr(i,l) - LogCorr(j,l) \right| + (N-2) \times (1 - LogCorr(i,j))$$

В качестве алгоритма кластеризации для заданной полуметрики $r(i,j)$ и фиксированного числа кластеров используется иерархическая группировка, в которой расстояние между кластерами определялось согласно функции:

$$r(K_p, K_q) = \max_{i \in K_p, j \in K_q} (r(i,j))$$

После нахождения $1 \le t \le n$ кластеров в сокращенную подсистему признаков включаются наиболее информативные признаки, по одному из каждого кластера. Таким образом находятся подсистемы из t наиболее информативных и некоррелированных признаков.

В третьей главе («Модели распознавания кореферентной связи») приводится постановка общей задачи для определения кореферентной связи между анафо́ром и антецедентом, модель **MB** распознавания кореференции, а также модель **DSE** распознавания кореференции.

Постановка общей задачи для определения кореферентной связи между анафо́ром и антецедентом, модель **MB** *распознавания кореференции*

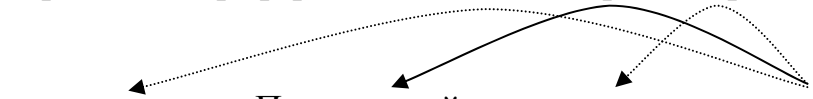
Общая задача определения кореферентной связи между анафо́ром и антецедентом состоит в следующем. Задан некоторый дискурс и в нём выделено местоимение третьего лица. Необходимо определить кореферентную связь анафо́ра с одним из гипотетических антецедентов, стоящих ранее по тексту. Сведем данную задачу к решению задач распознавания по прецедентам, в которых исходная информация задается анализаторами, обрабатывающими размеченные и неразмеченные тексты.

Пусть некоторому анафо́ру *Анф* соответствует совокупность гипотетических антецедентов $ГА_1$, $ГА_2$, … , $ГА_l$, один из которых является истинным ($\forall Анф \exists ГА_i : coref(Анф, ГА_i) = 1$, $i \in 1,\dots,l$, $\forall j \neq i, coref(Анф, ГА_j) = 0$). В качестве объектов S будем рассматривать совокупности <*Анф, ГА*$_1$, *ГА*$_2$, … , *ГА*$_l$> (S = <*Анф, ГА*$_1$, *ГА*$_2$, … , *ГА*$_l$>). Пусть паре <*Анф, ГА*$_i$> соответствует

признаковое описание $x(Анф,\ ГА_i) = (c_1(Анф,\ i),\ c_2(Анф,\ i),\ \dots,\ c_n(Анф,\ i))$, $c_t(Анф, i)$ – значение признака номер t относительно пары $<Анф,\ ГА_i>$. Признаковым описанием $I(S)$ объекта S будем считать вектор строку $I(S)=(x(Анф,\ ГА_1),\ x(Анф,\ ГА_2),\ \dots,\ x(Анф,\ ГА_l))$ размерности $n \times l$. В качестве множества M рассматриваем множество всех совокупностей $S=<Анф,\ ГА_1,\ ГА_2,\ \dots,\ ГА_l>$, допустимых в русском языке. Разбиение на классы определяется реальным антецедентом соответствующего анафо́ра.

Пусть задан некоторый размеченный корпус текстов. Выделяется множество всех анафо́ров и соответствующих им антецедентов, т.е. находятся множества совокупностей $M_i=\{S=<Анф,\ ГА_1,\ ГА_2,\ \dots,\ ГА_i>\}$, $i=1,2,\dots,L$. Признаковые описания объектов множества M_l и образуют таблицу обучения $T_{|M_l|,n\times l,l}$ для задачи распознавания с l классами. По данным обучающим таблицам строятся стандартные алгоритмы распознавания A_l с l классами, $l=2,3,\dots,L$.

Пример 3. Схема референциального выбора для трёх кандидатов



$l=3$

Вектор признаков для примера 3 в таблице $T_{|M_3|,n\times 3,3}$ из второго класса:

Анф=он, *ГА*$_3$=стиль	*Анф*=он, *ГА*$_2$=Паустовский	*Анф*=он, *ГА*$_1$=талант
94 признака	94 признака	94 признака

→ класс = II

Пусть задан некоторый анафо́р *Анф*. Определяется соответствующее ему число l гипотетических антецедентов. Кореферентная связь анафо́ра с одним из них устанавливается в результате решения задачи распознавания алгоритмом A_l. Данную модель распознавания кореференции обозначим **MB** (общая модель).

Модель **DSE** *распознавания кореференции*

В главе также описана модель распознавания кореференции, основанная на решении специальной дихотомической задачи распознавания в пространстве признаковых описаний $I(S_i) = (x_1(S_i), x_2(S_i), \dots, x_n(S_i))$ и задач распознавания в пространстве оценок (модель **DSE**).

В качестве множества $M=\{S\}$ допустимых объектов S возьмем допустимые наборы $<Анф,\ ГА_i>$, где $ГА_i$ – произвольный гипотетический антецедент для анафо́ра *Анф*. Первый класс K_1 образуют наборы, в которых $ГА_i$ – реальный антецедент, $i=1,2,\dots,L$ ($K_1=\{S\}$: $coref(Анф,\ ГА_i)=1$), $K_2 = M\backslash K_1$). Признаковым описанием S является $x(Анф, ГА_i) = (c_1(Анф, i),\ c_2(Анф, i),\ \dots,\ c_n(Анф, i))$.

Три вектора признаков для примера 3:

Анф=он, *ГА*$_3$=стиль		→ класс = II
Анф=он, *ГА*$_2$=Паустовский	✓	→ класс = I
Анф=он, *ГА*$_1$=талант		→ класс = II
94 признака	выбор эксперта	

Таким образом, совокупности таблиц $T_{|M_l|,n\times l,l}, l = 2,3,\dots,L,$ можно поставить в соответствие таблицу $T_{m^*,n,2}$, где $m^* = \sum_{i=1}^{L} |M_i| \times i$ (для простоты будем обозначать далее данную информацию как I_0). При этом число представителей первого класса в таблице $T_{m^*,n,2}$ равно $\sum_{i=1}^{L} |M_i|$. Задачу отнесения S по начальной информации $T_{m^*,n,2}$ к одному из двух классов обозначим как задача $\mathbf{Z}_0$. Отнесение S в один из классов соответствует ответу на вопрос, имеется или нет кореференция анафо́ра с соответствующим *ГА*. Данный ответ не может считаться окончательным, поскольку в задаче $\mathbf{Z}_0$ рассматривается вопрос связи анафо́ра с каждым из *ГА* независимо друг от друга. Таким образом, после решения задачи $\mathbf{Z}_0$ требуется разработка уточнённого алгоритма кореференции, учитывающего информацию о кореференции анафо́ра с группой *ГА* в дискурсе.

Пусть построен некоторый стандартный алгоритм распознавания $A=R\text{x}r$ для решения задачи $\mathbf{Z}_0$, где R – распознающий оператор, r – решающее правило. $R(I_0, x(Анф, ГА_t)) = (a_t^1, a_t^2)$, где a_t^1, a_t^2 – оценки, вычисляемые распознающим оператором за первый и второй классы, соответственно.

Сформулируем новую задачу распознавания $\mathbf{Z}_0{}^*$. Множество допустимых объектов $M^* = \{S^*\}$ формируется по результатам применения распознающего оператора к признаковым описаниям x(*Анф, ГА*$_1$), x(*Анф, ГА*$_2$), … , x(*Анф, ГА*$_l$) (где l – число ГА некоторого *Анф*), соответствующих всем ГА анафо́ров *Анф* – числовые векторы-строки $I(S^*) \equiv S^* = (a_1^1, a_1^2, a_2^1, a_2^2, \dots, a_l^1, a_l^2)$. Тогда $M^* = \bigcup_{i=2}^{L} M_i,\ M_i \bigcap M_j = \varnothing, i, j = 1,2,\dots,L, i \neq j.$ Здесь M_i^* – совокупность объектов $S^* = (a_1^1, a_1^2, a_2^1, a_2^2, \dots, a_i^1, a_i^2)$. Разбиение на классы множеств M_i^* задается порядковыми номерами реальных антецедентов, соответствующих исходному для S^* элементу $S =$<*Анф, ГА*$_1$, *ГА*$_2$, … , *ГА*$_i$>. Задачу распознавания объектов из M_i^* обозначим как $\mathbf{Z}_0^i{}^*$ а соответствующую начальную (обучающую информацию) как $I_0^i{}^*$. Алгоритмы решения задачи $\mathbf{Z}_0^i{}^*$ обозначим A^* и будем искать их в виде стандартных распознающих алгоритмов $A^* = R^* \text{x} r^*$. Рассматривались два варианта их построения.

1. Алгоритмы решения задач $\mathbf{Z}_0^i*$ с решающим правилом максимума оценок:

$$R*(I_0^i*,S)=(a_1^1,a_2^1,...,a_i^1),\ r*(a_1^1,a_2^1,...,a_i^1)=\begin{cases}t, & a_t^1>a_j^1, j=1,2,...,i, j\neq t,\\ 0, & \text{иначе}\end{cases}$$

2. Алгоритмы решения задач $\mathbf{Z}_0^i*$ как стандартные алгоритмы распознавания в пространствах оценок $(a_1^1,a_1^2,a_2^1,a_2^2,...,a_l^1,a_l^2)$.

В данном случае для заданной задачи $\mathbf{Z}_0^i*$ рассматриваются стандартные методы построения распознающих операторов и решающих правил для различных базисных моделей: алгоритмы вычисления оценок, голосование по тупиковым тестам, метод опорных векторов и др.

Анф=он, *ГА*$_3$=стиль	*Анф*=он, *ГА*$_2$=Паустовский	*Анф*=он, *ГА*$_1$=талант	→ класс = II
α_1^1 α_1^2	α_2^1 α_2^2	α_3^1 α_3^2	

В четвертой главе («Программный комплекс распознавания кореферентных связей и результаты практических применений») приводится технологическая карта программного комплекса распознавания кореферентных связей, результаты практических применений моделей распознавания кореференции **MB** и **DSE** и результаты поиска оптимального признакового пространства.

Для решения задачи распознавания был создан программный комплекс распознавания кореферентных связей, включающий программы подготовки и обработки ЕЯ-текстов, сервисные и вспомогательные программы, программную систему распознавания по прецедентам «РАСПОЗНАВАНИЕ». Программы созданы в соответствии с общей алгоритмической моделью подготовки и обработки ЕЯ-текстов. Модель направлена на нахождение статистических данных, описывающих процесс референциального выбора, и влияющих факторов в целях дальнейшего применения ММРО (Схема, с.20).

На базе новостных лент двух информационных агентств сформировано две обучающие выборки S^1_{ini} (2167 объект класса K_1 и 11238 объектов класса K_2) и S^2_{ini} (127 объектов класса K_1 и 1239 объектов класса K_2), где классы определяются согласно модели **DSE**.

При формировании обучающих выборок использовалось q=2 – число предложений, в которых производится отбор гипотетических антецедентов.

Для обучения использовалась подмножество объектов S^1_{ini}. С помощью генератора случайных чисел выборка S^1_{ini} была разбита на две подвыборки:

– обучающую выборку S^1_{tr}, по 1000 объектов из K_1 и K_2;

– контрольную выборку S^1_{contr}, включающую не вошедшие в S^1_{tr} объекты S^1_{ini} (1186 объектов из класса K_1 и 10219 объектов из класса K_2).

Для распознавания выбирались 11, 42, 61 и 84 лучших признаков по функционалу F.

Для решения задачи распознавания использовались методы, вошедшие в систему интеллектуального анализа данных «РАСПОЗНАВАНИЕ» [Журавлев, Рязанов, Сенько, 2006]:

– метод q-ближайших соседей, версия алгоритма с поиском оптимального числа ближайших соседей по обучающей выборке в режиме скользящего контроля;

– линейный дискриминант Фишера;

– линейная машина;

– метод ABO (алгоритмы вычисления оценок), вариант метода с голосованием по всевозможным опорным множествам;

– метод опорных векторов (SVM), вариант метода с гауссианой размера 6,0 в качестве потенциальной функции;

– логические закономерности;

– статистически взвешенные синдромы (СВС), «быстрый» вариант метода с разбиениями интервалов допустимых значений признаков одной точкой и без дополнительного отбора признаков.

Определение кореферентного антецедента в модели **MB**

В таблице 1 приведены результаты распознавания в модели **MB**, где в качестве стандартного алгоритма распознавания A_1 использовался метод опорных векторов.

Табл. 1.

обучающая выборка	точность распознавания, %	число объектов	число признаков задачи
M_2	84,2	292	188
M_3	74,1	352	282
M_4	63,5	307	376
M_5	51,6	250	470
M_6	55,5	232	564
M_7	49,5	186	658
	Совокупная точность распознавания:		64,95%

Учитывая наличие 168 объектов, для которых в качестве альтернативы был всего лишь один кандидат, совокупная точность алгоритма референциального выбора, построенного на базе модели распознавания, составила **68,24%**. Представляется, что малая точность распознавания обусловлена большим числом признаков, увеличивающимся с ростом числа гипотетических антецедентов, а также сравнительно малым числом объектов в каждой из выборок.

Определение кореферентного антецедента в модели **DSE**

В табл. 2 приведены результаты распознавания контрольных выборок S^1_{contr} и S^2_{ini} алгоритмами решения задач $\mathbf{Z}^i_0*$ с решающим правилом максимума оценок. Столбцы (1) – распознавание связей для пар ($Анф_x$, $ГА_i$) по модели DSE (таблица $T_{m*,n,2}$). Столбцы (2) – оценки результата работы решающего правила по числу верно установленных кореферентных связей в дискурсе для анафора по модели DSE (стандартный распознающий алгоритм A^* с решающим правилом максимума оценки).

Табл. 2. Точность распознавания модели с правилом максимума оценок

	распознавание объектов из выборки S^1_{contr}		распознавание объектов из выборки S^2_{ini}	
метод	(1), %	(2), %	(1), %	(2), %
SVM	**79,9**	78,2	74,8	**77,9**
CBC	79,8	68,0	**78,1**	60,6
Линейная машина	81,5	**79,8**	72,2	70,9

Результаты решения задач $\mathbf{Z}^i_0*$ с применением стандартных алгоритмов распознавания в пространствах оценок представлены в таблице 3, где приводится точность обучения и распознавания таблиц методом опорных векторов.

Табл. 3. Точность распознавания задач $\mathbf{Z}^i_0*$ методом опорных векторов

задача	число классов задачи	точность распознавания, %	число объектов	число признаков задачи
$\mathbf{Z}^2_0*$	l=2	92,0	292	4
$\mathbf{Z}^3_0*$	l=3	86,2	352	6
$\mathbf{Z}^4_0*$	l=4	82,4	307	8
$\mathbf{Z}^5_0*$	l=5	79,6	250	10
$\mathbf{Z}^6_0*$	l=6	73,0	232	12
$\mathbf{Z}^7_0*$	l=7	66,0	186	14
	Совокупная точность распознавания:			81,29%

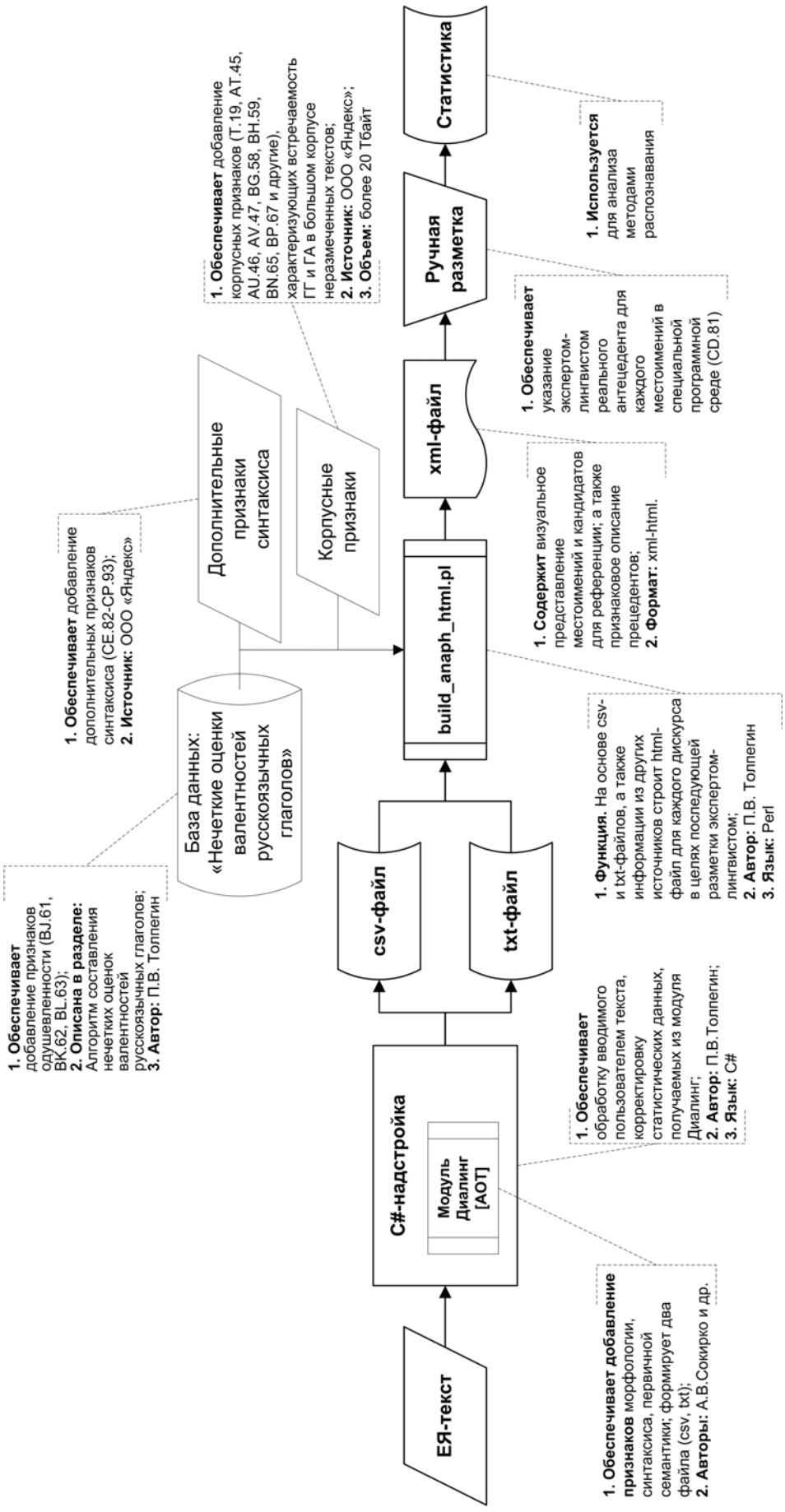


Схема. Технологический процесс подготовки и разметки ЕЯ-текстов

Учитывая наличие 168 объектов, для которых в качестве альтернативы был всего лишь один кандидат, который выбирался исходя из условий построения системы, совокупная точность алгоритма референциального выбора, построенной на базе модели распознавания, составила **83,05%**. Результаты решения задач $\mathbf{Z}_0^i$* с применением метода опорных векторов в пространствах оценок показало бóльшую точность (83,05%), чем решающее правило максимума оценок (79,8%).

Найдены оптимальные признаковые подпространства при использовании методов достоверных статистических разбиений, тупиковых тестов и логических корреляций. Оценки точности распознавания для различных признаковых подпространств приведены на рис.1. Следует отметить, что в оптимальные признаковые подпространства входила значительная часть корпусных признаков (BH59, BG58, CB79 и другие), подпространства, полученные с помощью различных подходов, имели существенное пересечение.

На рис.1 приводится оценка точности распознавания в скользящем контроле S^1_{tr} – подвыборки обучающей выборки S^1_{ini}, содержащей в каждом классе по 1000 случайным образом выбранных объектов.

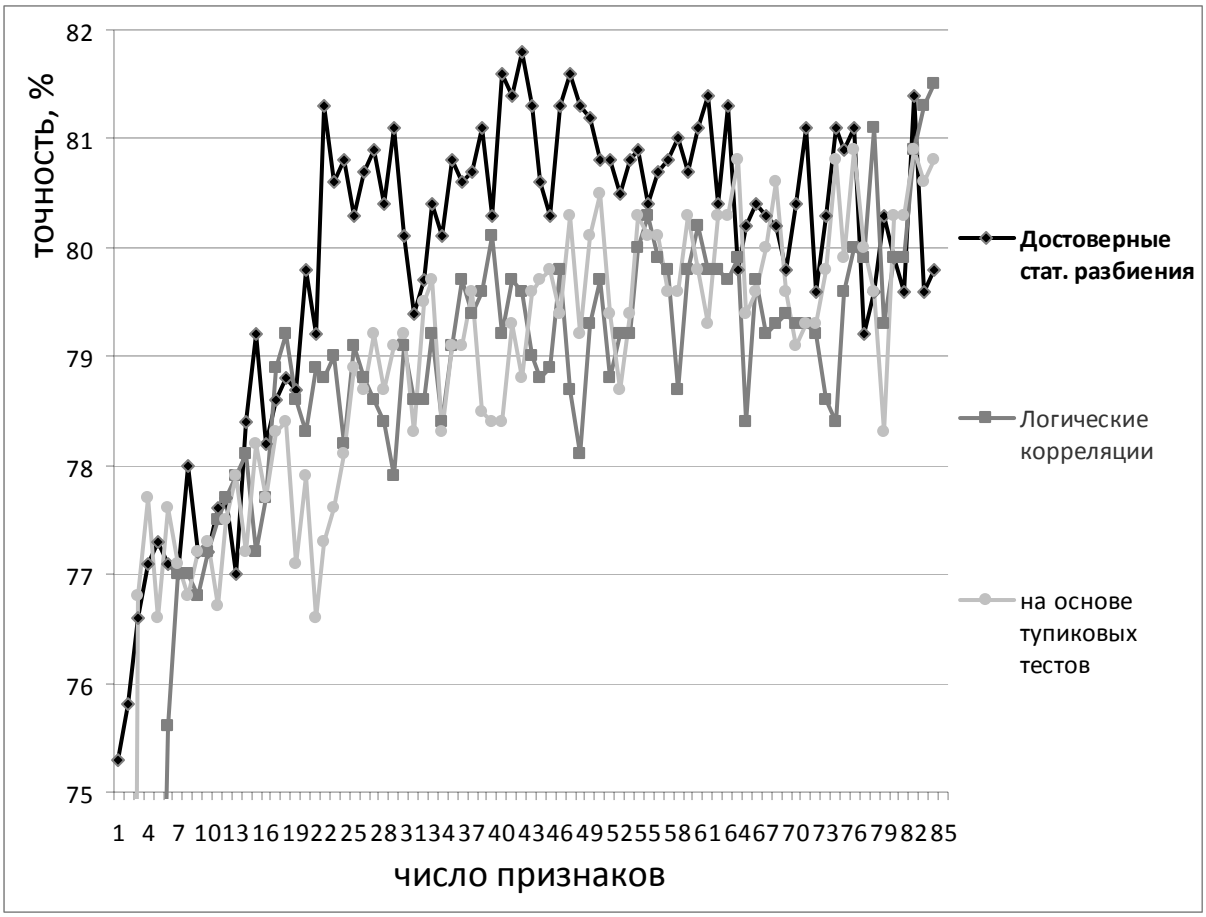


Рис. 1. Эффективность выбора оптимального признакового пространства

Исследованы изменения в точности распознавания выборки S^1_{tr} при ранжировании числа признаков тремя различными способами: на базе метода достоверных статистических разбиений, на основе тупиковых тестов, на основе логических корреляций. Результаты расчетов показали предпочтительность применения метода достоверных статистических разбиений, использование оптимальных признаковых подпространств из 20-25 признаков практически не снижало точность распознавания.

В заключении сформулированы основные теоретические и практические результаты, полученные в ходе работы над диссертацией.

Семантика остается в центре внимания для современных задач прикладной и математической лингвистики, в частности – в задаче автоматического разрешения местоименной ана́форы. В связи с труднодоступностью полноценных семантических словарей и иных ресурсов, в исследовании задействованы неразмеченные корпуса текстов больших объёмов, а также методы и алгоритмы по извлечению специфичной статистической информации в целях её использования в качестве системы признаков в задаче машинного обучения.

На предмет установления кореферентных связей между анафо́ром (местоимением) и антецедентом (существительным, причастием и проч.) вручную размечен информационно-новостной корпус (2186 дискурсов), на порядок превышающий объёмы текстов, на которых проводились опыты отечественными и зарубежными разработчиками.

Серия экспериментов по анализу сформированной статистической выборки показала точность в 83,05% (модель DSE, распознавание в пространстве оценок) и 77,9% (модель DSE, с решающим правилом максимума оценок) на независимых тестовых выборках, что приблизительно на 20% выше, чем аналогичный результат, полученный без использования новых введённых признаков.

По итогам анализа контекстов причина ошибок в точности обусловлена:

(1) ошибкой анализаторов в снятии морфологической и синтаксической омонимии – 9,45%;

(2) недостатком статистических данных, получаемых теориями фокуса и центрирования – 2,7%;

(3) другими ошибками – 4,8%.

Полнота (отношение числа анафорических местоимений третьего лица, для которых экспертом было успешно проведено связывание с реальным антецедентом, к общему числу анафорических местоимений третьего лица) составила 79,2% на тестовой выборке. Соответственно, 20,8% ошибок полноты работы системы вызваны:

(1) накапливающейся ошибкой в снятии омонимии используемыми анализаторами. Так, например, «*Полевой*», являясь реальным антецедентом, определялся не как имя собственное мужского рода, а как прилагательное женского рода. Следовательно, в рассматриваемом дискурсе невозможно

было осуществить связывание анафо́ра и антецедента. При анализе таких местоимений как *его*, *их* и др., которые имеют морфологическую омонимию, гипотетические антецеденты выбирались в соответствии с родом (и числом), идентифицированным анализатором, что также влияло на полноту – 16,45%;

(2) ограничением в q=2 предложения – как область поиска антецедента, т.е. реальный антецедент мог существенно предшествовать анафо́ру по тексту – 2,8%;

(3) ошибками иного рода – 1,55%.

Если на начальном опыте при анализе процента ошибок точности (38%) на обучающей выборке отмечалось влияние выразительных средств языка, то анализ значимости новых признаков показал, что введённые корпусные признаки занимают места с номерами 4, 5, 9, 10, 13, 16, 28, 33 в перечне из 94 позиций, ранжированном по убыванию функционала качества. Именно эти признаки, как выявил контекстный анализ, большей частью компенсируют нехватку семантики, логики и знаний о Мире, что является новым в развитии рассматриваемой проблемы.

Признаки, связанные с одушевлённостью, занимают менее высокие позиции: 14, 57, 68, 73. Симптоматично, что не столь высокая эффективность признаков одушевлённости как у корпусных признаков объясняется не таким частым появлением контекстов, в которых одушевлённость играет решающую роль. Гипотетические антецеденты могут быть одинаково одушевлённым или неодушевлёнными, однако после применения вышеуказанных признаков остается более одного антецедента «прогнозируемой» одушевлённости.

Вместе с тем, если корпусные признаки «строились» на объёме текста порядка 20 Тбайт, то оценки степени одушевлённости – на тексте около 80 Мбайт по причине значительной вычислительной сложности эксперимента. Различия в объёмах входных данных при построении мер оказали влияние на их полноту. При создании обучающей выборки S_{tr} три оценки одушевлённости «отказались» возвратить значение за отсутствием таковой в 15,8; 16,5 и 16,5% случаев, т.е. полнота мер №№ 1-3 составила 84,2; 83,5 и 83,5% соответственно.

Алгоритмы анализа, обеспечивающие формирование показателей фокуса и центра, не оценивались в силу незначительной полноты их применения. Функционирование указанных алгоритмов определяется точностью синтаксического анализатора, поэтому число случаев их применимости и безошибочного определения зависимых слов, оказались ощутимо малыми.

———

ОСНОВНЫЕ РЕЗУЛЬТАТЫ РАБОТЫ

1. Разработаны общая модель распознавания кореференции (МВ) и модель распознавания кореференции, основанная на решении специальной дихотомической задачи распознавания в пространстве признаковых описаний и задач распознавания оценок (DSE). Полнота и точность модели DSE составили 79,2% и 83,05% соответственно.
2. Предложены и программно реализованы алгоритмы, формирующие расширенное признаковое пространство в задаче разрешения местоименной ана́форы третьего лица для русскоязычных текстов:
    а. алгоритм составления оценок степени встречаемости одушевлённости для валентностей русскоязычных глаголов;
    б. алгоритм синтаксической деривации;
    в. алгоритм определения конфликтующих антецедентов;
    г. алгоритм некореферентности анафо́ра с гипотетическим антецедентом;
    д. алгоритм некореферентности местоимений;
    е. алгоритмы по формированию корпусных оценок встречаемости гипотетического антецедента и глагольной группы, управляющей анафо́ром.
3. Методами математического обучения исследована эффективность корпусных признаков (оценок встречаемости гипотетического антецедента и глагольной группы, управляющей анафо́ром) при принятии решения референциального выбора.
4. Тремя подходами получены результаты анализа признаковых систем и информативные системы признаков, исследованы системы признаков и информативных подмножеств признаков.
5. Создан комплекс программ для ЭВМ, обеспечивающих предобработку ЕЯ-текстов и вычисление значений признаков.

———

Список публикаций по теме диссертации

1. Толпегин П.В. Информационно-поисковая система своими руками. XXIX Международная молодежная научная конференция «Гагаринские чтения», т. 5, М.: ИЦ «МАТИ», 2003, с. 16-17
2. Толпегин П.В. Словоформы Русского Языка для Информационно-Поисковой Системы. Свидетельство Роспатента о регистрации базы данных № 2003620059, 2003
3. Толпегин П.В. Программа искусственной генерации словоформ английского языка. Свидетельство Роспатента о регистрации программы для ЭВМ № 2003610875, 2003
4. Толпегин П.В. Программа искусственной генерации словоформ русского языка. Свидетельство Роспатента о регистрации программы для ЭВМ № 2003610874, 2003
5. Толпегин П.В. Программа поиска и восстановления словоформ по базе данных. Свидетельство Роспатента о регистрации программы для ЭВМ № 2003610871, 2003
6. Толпегин П.В. Текстовый поиск по сходству. XXX Международная молодежная научная конференция «Гагаринские чтения», т. 5., М.: ИЦ «МАТИ», 2004, с. 62-63
7. Толпегин П.В. Технологические приемы построения текстовых информационно-поисковых систем. М.: Издательский центр «МАТИ», 2004, с. 1- 73
8. Толпегин П.В. Агентно-ориентированный подход к построению корпоративных систем безопасности с применением методов классификации и распознавания. XXXI Международная молодежная научная конференция «Гагаринские чтения», т. 4., М.: ИЦ «МАТИ», 2005, с. 38-39
9. Толпегин П.В. Машинное обучение в референциальном анализе русских естественно-языковых текстов. Международная молодежная научная конференция XXXII «Гагаринские чтения», т. 4., М.: «МАТИ», 2006, с. 48-49
10. Толпегин П.В., Ветров Д.П., Кропотов Д.А. Алгоритм автоматизированного разрешения анафоры местоимений третьего лица на основе методов машинного обучения. Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2006» (Бекасово, 31 мая – 4 июня 2006 г.) / Под ред. Н.И. Лауфер, А.С. Нариньяни, В.П. Селегея. – М.: Изд-во РГГУ, 2006, 648 с.: ил. с. 504-507
11. Толпегин П.В. Информационные технологии анализа русских естественно-языковых текстов. Часть I. // Информационные технологии. – 2006. – №8. – С. 41-50
12. Толпегин П.В. Информационные технологии анализа русских естественно-языковых текстов. Часть II. // Информационные технологии. – 2006. – №9. – С. 2-7
13. Толпегин П.В., Ветров Д.П., Кропотов Д.А. Прагматический анализ с применением подходов к автоматизированному созданию онтологической базы данных. Десятая национальная конференция по искусственному интеллекту с международным участием КИИ-06 (25-28 сентября 2006 г., Обнинск): Труды конференции. В 3-т. Т.2. – М.: Физматлит, 2006, с. 498 – 505
14. Толпегин П.В. Новые методы и алгоритмы автоматического разрешения референции местоимений третьего лица русскоязычных текстов. М.: КомКнига, 2006, – 88 с.
15. Толпегин П.В. Автоматизированная межклаузная референция в задаче когнитивного анализа текстов. Информационные технологии в бизнесе: Тезисы докладов научно-технической конференции студентов, аспирантов и молодых специалистов. – М.: Государственный университет – Высшая школа экономики. 2006, с. 115-118
16. Толпегин П.В. Роль корпусных ресурсов поисковых систем в формировании признакового пространства для разрешения местоименной анафоры. Материалы 7-ой международной конференции Информационное общество, интеллектуальная обработка информации, информационные технологии. 24-26 октября 2007 г. НТИ-2007 М.: ВИНИТИ РАН, с. 314-317
17. Толпегин П.В. Формирование нечетких мер валентностей русскоязычных глаголов. Отчет конкурса «Интернет-математика» ООО «Яндекс», 2007, [Электрон. документ]. (http://download.yandex.ru/IMAT2007/tolpegin.pdf)
18. Толпегин П.В. Разработка и реализация методов семантического и прагматического анализов ЕЯ-текстов русского языка (грант № 06-06-80464-а). ВЦ РАН. Москва. Информационный бюллетень РФФИ №14. М.: Наука, 2007