ЦЕНТРАЛИЗОВАННОЕ ПЛАНИРОВАНИЕ VS. СТИХИЯ РЫНКА: СРАВНЕНИЕ ЛЕКСИЧЕСКОГО И ЖАНРОВОГО РАЗНООБРАЗИЯ В НАЦИОНАЛЬНОМ КОРПУСЕ РУССКОГО ЯЗЫКА И ИНТЕРНЕТЕ

CENTRAL PLANNING VS. FREE MARKET: COMPARING THE DISTRIBUTION OF TOPICS AND GENRES IN THE RUSSIAN NATIONAL CORPUS AND INTERNET

Шаров С.А.
Университет Лидса, Великобритания (s.sharoff@leeds.ac.uk)

В этой работе проводится сопоставление традиционных больших корпусов, таких как Британский Национальный Корпус или Национальный Корпус Русского Языка, и корпусов, полуавтоматически извлеченных из Интернета. Первый способ предполагает ручное аннотирование выборки из Интернет-корпуса и сравнение ее параметров с традиционным корпусом. Второй способ использует статистические модели, которые дают оценку жанрового и лексического разнообразия на основе автоматической кластеризации.

1. Введение

Совершенно очевидно, что использование больших корпусов позволяет получить более точное описание многих языковых явлений. Однако их создание является весьма трудоемкой задачей: сбор тысяч текстов, снятие проблем с авторскими правами, приведение всех текстов в единую форму, балансировка корпуса по темам и жанрам отнимают много времени. Это является основной причиной того, что большие сбалансированные корпуса существуют для очень немногих языков. Флагманом здесь является 100-миллионный Британский национальный корпус (БНК), созданный в первой половине 1990-х годов и призванный отразить различные функциональные сферы использования британского варианта английского языка [1]. Близок к завершению проект создания Национального корпуса русского языка (НКРЯ), во многих отношениях аналогичного БНК [2, 3]. Но такие корпуса отсутствуют даже для некоторых мировых языков, таких как китайский или французский, не говоря уже о многих других. А другие существующие корпуса, например, для немецкого [4], не имеют подробной метатекстовой разметки. Наконец, очень немногие большие корпуса помимо БНК доступны в виде исходных текстов. С другой стороны, естественным источником данных для лингвистических исследований является Интернет. Одним из стандартных способов построения корпусов на основе Интернета является использование коллекций текстов, таких как Проект Гутенберг http://www.gutenberg.org/ или Библиотека Максима Мошкова http://www.lib.ru/. Еще один способ заключается в использовании текстов новостных сайтов, например, Синьхуа http://www.xinhuanet.com/ или Страна.Ру. Однако, в сравнении с корпусами типа БНК или НКРЯ такие коллекции не могут считаться представительными, поскольку они не отражают разнообразие тем и функциональных стилей. Для лингвистов самым распространенным способом работы с Интернетом остается составление запросов к поисковой машине и интерпретация результатов либо по числу найденных страниц, либо по первым возвращенным ссылкам. В английском такая методология получила название Googleology [5], для русского более подходящим названием может стать Яндексология.

Аргументы за создание традиционных корпусов таких как БНК или НКРЯ в отличие от использования Интернета можно кратко изложить следующим образом. Во-первых, содержимое Интернета не является сбалансированным корпусом: набор текстов в Интернете отражает предпочтения и интересы его активных пользователей, что может существенно исказить картину описываемых явлений. Во-вторых, поисковые машины ориентированы на информационный поиск: в их запросах не могут быть использованы лингвистические критерии (например, грамматические признаки), а результат поиска отсортирован по информационной релевантности найденных страниц. Слова запроса могут также встречаться не в самом тексте найденных страниц, а в их названиях, списке ключевых слов или даже на других страницах, ссылающихся на данную. В-третьих, многие лингвистические рассуждения, использующие Интернет, основаны на сравнении количества страниц, найденных для

нескольких запросов, но этим данным не всегда можно доверять: Жан Верони отметил, что Гугл находит в два раза больше страниц по запросу Chirac, чем по Chirac OR Sarkozy (http://aixtal.blogspot.com/2005/02/web-googles-missing-pages-mystery.html). Наконец, в-четвертых, невозможно провести статистическую оценку результатов: невозможно найти слова, часто встречающиеся в контексте запроса (коллокации), или хотя бы оценить какую пропорцию общего множества составляет запрос. В результате считается, что единственно правильным способом эмпирического описания языковых явлений является использование не Интернета, а традиционных представительных корпусов [6].

Легко видеть, что за исключением первого аргумента все остальные относятся к уровню интерфейса поисковой машины. В связи с этим Адам Килгаррифф предложил идею создания поисковой машины для лингвистов, которая должна обходить весь Интернет как обычная поисковая машина, но будет выдавать результаты в формате пригодном для лингвистов [7]. Эта идея не была до сих пор реализована в полной мере, предварительный эксперимент описан в [8], но ее воплощение не снимает первого вопроса: какие типы текстов находятся в Интернете и как сильно они отличаются от традиционных корпусов.

В этой статье я попытаюсь построить статистический портрет Интернет-корпусов и сравнить его с традиционными корпусами. Сначала будет представлен метод создания Интернет-корпуса на основе случайных запросов к поисковым машинам (этот подход позволяет избежать необходимости построения механизма обхода Интернета). Затем темы и жанры таких корпусов будут сопоставлены с традиционными корпусами (БНК и тестовой версией НКРЯ). Этот метод сравнения корпусов можно описать как приближение Интернет-корпусов к традиционным, баланс тем и жанров в которых известен. В заключение будут приведены результаты кластерного анализа, что позволит статистически оценить их внутреннюю структуру независимо от их традиционных эквивалентов. Статистические модели были построены с помощью систем CLUTO [9] и Weka [10].

2. Создание интернет-корпусов

Предлагаемый механизм создания корпусов использует четыре шага:

1. выбор слов (создание списка из примерно 500 слов частых в языке X);
2. порождение запросов (создание списка из 5-8 тысяч запросов, каждый из которых состоит из трех-четырех слов, полученных на шаге 1)
3. получение Интернет-страниц (сохраняются десять-двадцать первых страниц, возвращаемых по исковой машиной для данного запроса)
4. пост-обработка (разрешение проблем с кодировкой, снятием навигации, удалением дубликатов)

На первом шаге можно использовать термины, задающие некоторую предметную область, чтобы получить специализированный корпус. Необходимость использования нескольких слов в запросе связана с необходимостью получить относительно длинные связные тексты и уменьшить количество шума, т.е. страниц нерелевантных для лингвистического анализа, например, прайс-листов, таблиц, списков ссылок и т.п. На основе экспериментов было установлено, что три-четыре общеупотребительных слова в запросе (например, событие, работа, поднимать) в большинстве случаев приводят к получению связных текстов разнообразной тематики. Использование более длинных запросов увеличивает требования на длину текста и существенно уменьшает количество текстов, выдаваемых поисковой машиной.

На последнем шаге после загрузки страниц необходимо выполнить несколько технических операций, таких как унификация кодировок (в зависимости от языка можно использовать либо функцию guess-encoding в Перле, либо программу Enca), удаление средств навигации внутри страницах и дубликатов между страницами. Стандартным алгоритмом для удаления дубликатов и почти одинаковых страниц (например, таких как газетная страница и ее версия для печати) является использование шинглов [10]. Средства навигации на Интернет страницах существенно искажают статистические параметры корпуса, повышая частоту таких выражений как Вернуться назад или Отправить комментарий. Для их удаления на произвольных страницах (в ситуации, когда мы не знаем заранее набор используемых навигационных выражений) используется эвристика Финна: средства навигации обычно компактно расположены на странице и могут быть найдены поиском областей с максимальной плотностью тэгов [11].

Эта последовательность шагов в результате позволяет получить большой корпус (100-200 миллионов) на любом языке, распространенном в Интернете. На данный момент на странице http: //corpus.leeds.ac.uk/list.html доступны корпуса английского, испанского, китайского, польского, португальского, финского, фарси, японского и некоторых других языков. Более подробно эта методика описана в [12]. В описании ниже мы будем ссылаться на Интернет-корпуса, полученные по этой методике, с помощью буквы I, за которой следует код языка, например, I-EN, I-DE и I-RU для английского, немецкого и русского.

3. Представительность Интернет-корпусов

Успех в создании Интернет корпусов не снимает проблемы с их представительностью. Если, как это часто утверждается, Интернет состоит главным образом из порнографии и спама, то и корпус, созданный на его основе будет представителен только для таких текстов. Для того, чтобы оценить тематическое и жанровое разнообразие Интернет-корпусов в сравнении с представительными корпусами, мы должны создать систему категорий, одинаковых для представительных корпусов и Интернета, сделать случайную выборку статистически достоверного набора страниц из последних и сравнить их баланс. Это нетривиальная задача, поскольку наборы категорий, используемых в метатекстовой разметке представительных корпусов, несовместимы между собой, и страницы Интернета не дают достаточной информации для определения значений многих категорий такой разметки, таких как возраст автора или количество предполагаемых читателей.

Система классификации текстов в БНК включает несколько параметров [1]. Так для определения тематики текстов используется девять категорий, включающих художественную литературу, естественные, прикладные и гуманитарные науки, религию и философию и т.п. Устные тексты в БНК не имеют тематической классификации. Тестовая версия НКРЯ также классифицирует тексты по нескольким параметрам [13], включая тему текста, которая кодируется списком из 52 категорий, имеющих разную степень подробности: от экономика или внутренняя политика до безопасность, путешествия или вооруженные конфликты; многие малые категории представлены всего одним-двумя текстами, а отнесение текстов к ним зависит от субъективных предпочтений кодировщика: тема сообщения о вызрыве в Чечне – это политика, безопасность или вооруженные конфликты? Художественные тексты в НКРЯ тематической классификации не имеют (в сущности это 53-я тематическая категория).

Для Интернет-корпусов была использована классификация, которая является дальнейшим упрощением классификации БНК и которая может быть использована для обобщенной классификации тем в НКРЯ, что дает возможность сравнить эти корпуса и между собой и с Интернетом:

- natsci – естественные науки (математика, биология, физика)
- appsci – технологии и прикладные науки (включая медицину)
- socsci – гуманитарные науки (эта категория включает религию и философию)
- politics – политика
- commerce – бизнес
- arts – искусство
- leisure – развлечения и зрелища (спорт, путешествия, мода и т.п.)
- life – общая категория для текстов без четко выраженной тематики, таких как художественная литература, анекдоты, чат на сайтах знакомств и т.п.

Ситуация с жанровой классификацией в меньшей степени поддается унификации. В НКРЯ тексты делятся на 9 групп по сфере применения: бытовая, официально-деловая, производственная, публицистика, публичная речь, реклама, учебно-научная, художественная и церковно-богословская. В дополнение к этому используется открытый список типов текстов, например, интервью, отзыв, закон (существующий список содержит более 100 типов). Эксперимент по классификации русскоязычного Интернета [14] использовал систему из пяти категорий стилей: официальный, академический, публицистика, литература и повседневный стили. Для БНК Дэвид Ли предложил классификацию [15], состоящую из 70 жанров, например, таких как гуманитарные научные тексты, популярные тексты по медицине или репортажи в общенациональных серьезных газетах (broadsheets). В англоязычных текстологических исследованиях [16, 17] часто используется классификация, состоящая из четырех категорий: Descriptive-Narrative (описательные тексты), Explicatory-Informational (информативные тексты), Argumentative-Persuasive (аргументативные тексты) и Instructional (инструктивные тексты). Однако, эта классификация не дает четких определений для категорий, например, научная статья – это описательный, информативный или аргументативный текст? С другой стороны, перечисление жанров открытым списком из 70-100 категорий не дает возможности разумного выбора для пользователя.

Для кодирования Интернет-корпусов была использована новая классификация, которая была построена на основе рекомендаций Джона Синклера [18] для оценки цели создания текста:

- discussion – тексты, обсуждающие тему, например, газетные или научные статьи (эта категория вбирает некоторые виды описательных и аргументативных текстов).
- recommendation – тексты, рекомендующие некоторый способ действий (что соответствует другим видам аргументативных текстов, таким как реклама, но также влючает законы).
- recreation – тексты, предназначенные для развлекательного чтения (еще один вид описательных текстов), к ним относится художественная литература, а также популярные биографии, бульвар ные газеты и т.п.

- information — тексты, предназначенные для предоставления информации. В соответствии с [18] эта категория предназначена для энциклопедической информации, но к информационным текстам видимо необходимо относить и информационные сводки разного рода.
- instruction – инструктивные тексты. По мысли Синклера она включает в себя, как собственно инструкции, так и учебники.

		БНК	I-EN	НКРЯ	I-RU	I-DE
Жанр	Discussion	49%	45%	-	47%	45%
	Information	1%	11%	-	4%	25%
	Advert	7%	34%	-	35%	21%
	Instruction	4%	6%	-	3%	5%
	Recreation	39%	4%	-	11%	4%
Тема	Life	27%	14%	51%	25%	12%
	Politics	19%	12%	18%	10%	21%
	Business	8%	13%	3%	7%	5%
	Natsci	4%	3%	2%	3%	1%
	Appsci	7%	29%	3%	19%	18%
	Socsci	17%	16%	16%	5%	8%
	Arts	7%	2%	6%	2%	4%
	Leisure	11%	11%	1%	26%	31%

Таблица 1. Сравнение жанрового и тематического состава корпусов

a	b	c	d	e	<- classified as
195	0	0	3	10	a = discussion
1	23	3	3	5	b = information
1	1	13	0	0	c = instruction
1	1	1	50	1	d = recommendation
3	0	1	0	509	e = recreation

Таблица 2. Качество распознавания жанров БНК в обучающей выборке

В таблице 1 приводятся результаты сравнения корпусов по жанрам и темам. Как видим Интернет корпуса не слишком отличаются от традиционных корпусов как по представленным жанрам, так и по темам. Наиболее существенные отличия касаются количества текстов, предназначенных для развлекательного чтения (категория recreation). За исключением русского языка такие тексты в Интернете представлены мало в связи с ограничениями на авторские права. Поскольку тексты художественной литературы тематически классифицируются как life, категория life также меньше представлена в Интернет-корпусах. Другая тематическая категория недопредставленная в Интернет-корпусах – гуманитарные науки и искусства. С другой стороны технические тексты представлены в Интернете в гораздо большем количестве по сравнению с традиционными корпусами.

Для русского и немецкого Интернет-корпусов оценка проводилась на основе выборки из 200 документов.

Но для английского можно получить более полную статистическую оценку. Хотя в БНК не проводится строгого различия по целям создания текстов, такие категории можно выделить для некоторых жанров. Например, в БНК отсутствует класс recommendation, но выделен жанр рекламных текстов, к классу recreation относятся тексты из художественной литературы, популярных биографий и бульварной прессы, которые также описаны в БНК. Это позволяет на основе подмножества из 838текстов создать статистическую модель, которую можно применить ко всему БНК и получить, таким образом, сравнимую оценку на основе большего корпуса. Статистическая модель была построена на основе частотной информации о последовательностях из трех частеречных кодов [17] и системы Weka с помощью SVM [19], что обеспечило 96% точности распознавания жанров (Таблица 2). Строки соответствуют исходным категориям, столбцы – результатам классификации случайным 10-кратным делением на обучающую и тестовую выборки (10-fold cross-validation), например, из 514 примеров развлекательных текстов 4 были классифицированы как дискуссионные.

4. Кластеризация

В предыдущем разделе мы сопоставили структуру традиционных корпусов и корпусов, полученные из Интернета. Но нет гарантии того, что существующие классификации достаточно хорошо отражают тематическую и жанровую специфику Интернета. Кроме того, применение машинного обучения ограничено необходимостью использовать большое количество текстов для тренировки. По этой причине стоит попробовать автоматическую кластеризацию Интернет корпусов.

В этом случае нам придется решить три проблемы. Во-первых, необходимо заранее задать количество кластеров: алгоритм кластеризации найдет любое заданное количество кластеров и существующие методы оценки качества кластеризации далеки от совершенства. Здесь следует исходить из того, что целью кластеризации является сравнение с существующими классификациями, поэтому заданное число кластеров должно слегка превосходить число классов, используемых в традиционных моделях. Это позволит с одной стороны обнаружить классы, которые стоит разбивать на отдельные категории, с другой стороны, это позволит объединить некоторые классы, между которыми не находится существенных различий. В этом эксперименте был использован алгоритм повторного разбиения (Repeated Bisections) системы CLUTOВ итерациях по 8-15 кластерам тем и 5-9 кластерам жанров. После этого были использованы результаты, которые поддавались лучшей интерпретации: 11 кластеров предметных областей и 7 кластеров жанров.

Во-вторых, кластеризация производится на основе признаков (features), описывающих каждый документ. Известно, что темы выделяются на основе ключевых слов, но точный набор ключевых слов заранее неизвестен, поэтому этот список был порожден автоматически. Сначала были выделены ключевые слова для каждого документа корпуса. После этого список всех ключевых слов был отсортирован по количеству документов, для которых это слово было ключевым. В качестве признаков были использованы 2000 слов, которые чаще оказывались ключевыми словами отдельных документов. Для кластеризации жанров была снова использована та же модель частеречных кодов (pos trigrams), которая показала свою силу при классификации жанров.

В-третьих, довольно сложно сравнивать результаты кластеризации на разных корпусах. Если кластеризация НКРЯ породила одно разбиение, кластеризация I-RU – другое, а I-EN – третье, невозможно в точности сказать, что кластер 1 НКРЯ соответствует кластеру 2 I-RU и кластеру 3 в I-EN. Чтобы сократить количество возможных точек сравнения, были порождены кластеры только для Интернет-корпусов. Для каждого кластера были выявлены наиболее специфичные признаки, которые и были сопоставлены с соответствующими списками в существующих классах БНК и НКРЯ. Списки ключевых слов отсортированы по значению логарифма правдоподобия (Log-Likelihood).

Результаты кластеризации тем представлены в Таблице 4. Каждый кластер был сопоставлен с классом по БНК и НКРЯ, с которым он пересекался по ключевым словам (для русского языка и БНК приведены леммы, для остальных корпуов словоформы). В результате видно, что традиционное разбиение на предметные области хорошо соответствует результатам машинной кластеризации, за исключением разбиения на естественные и прикладные науки (natsci и appsci), которое не нашло своего места среди кластеров, полученных для разных языков (соответствующие тексты были распылены между другими кластерами). С другой стороны, кластеризация породила отдельные кластеры для медицины и компьютеров. Вероятно это произошло в связи с наличием в этих областях большого количества специализированной терминологии, которая мало используется за их пределами. Кластеризация для английского языка также породила отдельный спортивный кластер, что привело к необходимости извлечь из БНК и НКРЯ ключевые слова по спортивной тематике и остальные зрелища и развлечения за пределами спорта (главным образом, путешествия). В Интернет-корпусах устойчиво выделяется также религиозная тематика (имеющая свою специфику в каждом из языков). Наконец, в каждом языке выделяются кластеры, типичные для данной культуры, например, регионы (Швейцария и Астрия для немецкого, страны бывшего СССР для русского, Тайвань для китайского), художественная литература для русского и немецкого, путешествия для русского и китайского.

Что касается набора ключевых слов для каждого кластера, для некоторых областей, например, политики и экономики, наблюдается существенное пересечение между кластерами и классами существующих корпусов. Этот эксперимент еще раз показывает ограниченность традиционных корпусов в покрытии специальных предметных областей. Статистические результаты радикально опровергают традиционную точку зрения, что Интернет состоит из порнографии и спама. Напротив, к интимной сфере относятся все ключевые слова, характеризующие медицинскую тематику в НКРЯ (что неудивительно, поскольку 90% таких текстов в тестовой версии НКРЯ покрыто книгой "1001 вопрос про ЭТО"). Аналогичным образом большая часть медицинских текстов БНК взята из журнала "Gastroenterology and hepatology", что и объясняет частоту ключевых слов этой тематики там. В Интернет-корпусах кластер медицинских текстов покрывает различные области медицины.

Вторая проблема традиционных корпусов касается их устаревания: современные слова в области компьютерных технологий гораздо лучше отражены в Интернет-корпусах (слово browser не употребляется в современном значении в БНК); ключевые слова БНК в области политики также отражают реалии конца 1980-х годов (слова Soviet, Thatcher, но не Blair, Putin). Упоминание имен конкретных людей для корпуса не так важно (корпус это не СМИ и не энциклопедия), но их деятельность повлекла за собой существенные изменения в словаре, например, в употреблении таких слов как target или вертикаль, для которых корпус должен предоставлять примеры современного употребления.

Entropy	recr	inst	disc	reco	info	Features
0.000	0	0	0	0	33	VBD DT JJ, VBD RB DT, DT VBZ JJ, VBD CD IN, CC PP PP
0.000	274	0	0	0	0	IN CD NN, NP NP (, NN MD RB, NP NN NN, CD NNS WP
0.027	137	0	0	1	0	IN CD NN, NP NP (, NN MD RB, JJ JJ NP, CD NP NN
0.596	10	0	27	3	2	MD VB RP, IN NNS WP, CD) CD, CD NNS WP, CC WDT VBD
0.132	86	0	5	0	0	VBD RB DT, VBD DT JJ, VBD CD IN, CC PP PP, NN NP NN
0.488	3	2	70	19	1	NP NP (, NP NN NN, CD NP NN, VBD DT JJ, PP JJ RB
0.573	3	13	106	31	2	VBZ JJ NNS, VBD DT JJ, DT VBZ JJ, PP IN NN, NN NNS IN

Таблица 3. Результаты кластеризации жанров БНК

Для жанров соответствие между кластерами и традиционными категориями построить гораздо сложнее, поскольку кластеры жанров описываются не ключевыми словами, а последовательностями частеречных кодов, например, для русского V A S (глагол прилагательное существительное) или для английского CC NN IN (союз существительное предлог), которые довольно сложно интерпретировать с жанровой точки зрения. Эксперимент по поиску типичных типов текстов проводил Дуглас Байбер [20], но его подход во-первых использует лингвистическую информацию о конструкциях (например, отсутствие that в конструкциях типа he said that), которые специфичны для конкретного языка, которые трудно определить автоматически и наличие которых уже само по себе ограничивает жанр (тот параметр, который мы хотим найти в результате). Во-вторых, Байбер не использует кластеризацию, а получает набор размерностей (Principal Component Analysis), например, 0.254(quote)+0.176(NP NP VBZ)-0.155(NN NN CC)-0.155(DT NN CC), которые в свою очередь нуждаются в интерпретации с жанровой точки зрения, например, приведенная выше размерность интерпретируется как informational vs. involved.

Поскольку кластеризация размеченного корпуса позволяет сравнить кластеры с известными классами, набор размеченных текстов БНК, использованный в эксперименте в Таблице 2 был подвергнут кластеризации, которая показала, что рекламные тексты легко идентифицируются как отдельный кластер, в то время как информативные и инструктивные тексты тяжело разделить на основе лишь частеречных кодов 3. В свою очередь две группы текстов: discussion и recreation разбиваются на несколько отдельных кластеров, что подтверждает успешность автоматической кластеризации, поскольку первая группа была получена объединением газетных и научных статей (существующих как отдельные категории в БНК), а вторая объединением текстов художественной литературы, популярных биографий и желтой прессы. Получившиеся кластеры лишь подтвердили разницу между стилистикой этих категорий.

5. Заключение

Изложенные эксперименты показывают, корпуса, извлеченные из Интернета, не слишком сильно отличаются от традиционных корпусов. Они покрывают примерно тот же спектр предметных областей и жанров текстов. В некоторых отношениях они даже лучше традиционных корпусов: их намного проще построить для нового языка, они лучше отражают тематическое разнообразие и ближе соответствуют современным темам. С другой стороны, в настоящее время Интернет значительно уступает традиционным корпусам в отношении разнообразия жанров. В наибольшей степени это относится к художественной прозе, которая, особенно для I-DE и I-EN гораздо менее представлена в Интернете по причине ограничений на авторские права. Результат кластеризации по ключевым словам (Таблица 4) позволяет также оценить примерное количество текстов определенной тематики в Интернете, например, политических новостей.

Дальнейшие направления исследований в области классификации текстов можно поделить на две группы: выделение категорий и признаков.

Категории Разнообразие подходов к описанию жанров, с одной стороны, отражает сложность самой проблемы жанровой классификации. С другой стороны, пользователи корпусов нуждаются в единой разумной классификации, которая позволит сравнивать корпуса и выделять их подмножества. Теоретики часто стремятся максимально подробно и точно описать каждый текст, выделяя в результате большое количество категорий, в качестве примера можно привести систему TEI [21] или библиотечные классификации. Жанров действительно много. Даже в пределах жанров академической коммуникации можно упомянуть научные и научно-популярные статьи, рецензии, монографии, приглашения на конференции, емейлы, списки рассылки, заявки на проекты, справки, протоколы заседаний и т.п. Кроме того, многие из них имеют аналоги и за пределами лишь только академической коммуникации. Выбор из списка нескольких сотен жанров мало реален. Но общепринятой типологии, содержащей обозримое количество категорий, пока не выработано. Например, кажется разумным выделять категорию инструктивных текстов, которая объединит и рецепты, и советы по ремонту автомобиля, но относятся ли к такой категории учебники и списки ЧаВО (FAQ)? Разумно выделять категорию канцелярских текстов, но в каком отношении они находятся к другим формам деловой коммуникации, например, к предложениям о сотрудничестве? Или последние в жанровом отношении подобны академическим заявкам на проекты?

Признаки После определения типологии текстов следующей задачей является выделение признаков для их автоматической классификации в Интернет-корпусах. Хорошо известно, что тему текста можно определить по ключевым словам, некоторые виды жанров можно определить по последовательностям частеречных кодов [17], но соответствие между жанрами и кодами далеко не однозначно. Кроме того, в связи с разнообразием видов жанров (информативные или инструктивные тексты vs. административная или учебно-научная сфера их применения) использование лишь одного набора признаков не кажется убедительным. Эксперимент в разделе 3 показал, что для успешной классификации требуется большая обучающая выборка, создание которой требует больших усилий. Одним из способов решения этой проблемы является выделение признаков, характеризующих найденные тексты данной категории и поиск текстов, максимально похожих на них в соответствии с этими признаками.

Список литературы

1. ASTON, GUY & Lou BURNARD: The BNC Handbook: Exploring the British National Corpus with SARA. Edinburgh University Press, Edinburgh, 1998.
2. ШАРОВ, С.А.: Представительный корпус русского языка в контексте мирового опыта НТИ, Серия 2, 5:8-19, 2003.
3. ПЛУНГЯН, В.А.: Зачем нужен Национальный корпус русского языка. // Национальный Корпус Русского Языка: 2003-2005, С. 6-20. Индрик, Москва, 2005.
4. GEYKEN, ALEXANDER: The DWDS Corpus: A Reference Corpus for the German Language of the 20th century. // FELLBAUM, C. (ред.): Idioms and Collocations: From Corpus to Electronic Lexical Resource Continuum, Birmingham, 2007.
5. KILGARRIFF, ADAM: Googleology is bad science. Computational Linguistics, 33(1), 2007.
6. KILGARRIFF, A.: The web as corpus. / / Proc. of Corpus Linguistics 2001, Lancaster, 2001.
7. BARONI, MARCO & ADAM KILGARRIFF: Large linguistically-processed Web corpora for multiple languages. // Companion Volume to Proc. of the European Association of Computational Linguistics, С. 87-90, Trento, 2006.
8. ZHAO, YING & GEORGE KARYPIS: Empirical and Theoretical Comparisons of Selected Criterion Functions for Document Clustering. Machine Learning, 55(3):311-331, 2004.
9. WITTEN, I.H. & E.. FRANK: Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, San Francisco, 2005.
10. BRODER, ANDREI Z., STEVEN C. GLASSMAN, MARK S. MANASSE & GEOFFREY ZWEIG: Syntactic Clustering of the Web. // Proc. Sixth International World-Wide Web Conference, 1997.
11. FINN, A., N. KUSHMERICK & B. SMYTH: Genre classification and domain transfer for information filtering. // Proc. European Colloquium on Information Retrieval Research, Glasgow, 2002.
12. SHAROFF, SERGE: Creating general-purpose corpora using automated search engine queries II BARONI, MARCO & SILVIA BERNARDINI (ред.): WaCky! Working papers on the Web as Corpus. Gedit, Bologna, 2006. http://wackybook.sslmit.unibo.it.
13. САВЧУК, С.О.: Метатекстовая разметка в Национальном корпусе русского языка II Национальный Корпус Русского Языка: 2003-2005, С. 62-88. Индрик, Москва, 2005.
14. BRASLAVSKI, PAVEL: Document style recognition using shallow statistical analysis. // ESSLLI 2004 Workshop on Combining shallow and deep processing for NLP, С. 1-9, 2004.

15. LEE, DAVID: Genres, registers, text types, domains, and styles: clarifying the concepts and navigating a path through the BNC jungle. Language Learning and Technology, 5(3):37-72, 2001.
16. SWALES, JOHN: Genre Analysis. English in academic and research settings. Cambridge University Press, Cambridge, 1990.
17. SANTINI, MARINA: Linguistic Facets for Genre and Text Type Identification: A Description of Linguistically-Motivated Features. ITRI-05-02, University of Brighton, 2005.
18. SINCLAIR, JOHN: Corpora for lexicography. // STERKENBERG, P. VAN (ред.): A Practical Guide to Lexicography, C. 167-178. Benjamins, Amsterdam, 2003.
19. PLATT, JOHN: Machines using Sequential Minimal Optimization. // SCHOELKOPF, B., C. BURGES & A. SMOLA (ред.): Advances in Kernel Methods – Support Vector Learning. MIT Press, 1998.
20. BIBER, DOUGLAS: Variations Across Speech and Writing. Cambridge University Press, 1988.
21. SPERBER-MCQEEN, CM. & L. BURNARD (ред.): TEI P4: Guidelines for Electronic Text Encoding and Interchange. Oxford, 4 , 2002.

Polit	BNC	government, minister, party, election, political, labour, president, soviet, military, police, parliament, thatcher, national, council, prime, public, report, state, former, country	19%
	I-EN	Japan, Minister, Zealand, Chinese, India, Government, WTO, Prime, trade, Senate, Foreign, South, Commonwealth, Senator, economic, Japanese, Australians, Sydney, Mr, Labor, Korea	6%
	I-EN	Labour, EU, European, Europe, war, US, political, Blair, military, British, UK, Minister, Iraqi, election, Britain, vote, international, party, London, aircraft, peace, defence, democracy	10%
	I-DE	Irak, Israel, militärisch, Hitler, Armee, Bush, jüdisch, russisch, Soldat, Iran, Kosovo, israelisch, Rußland, Nordkorea, irakisch, arabisch, Präsident, amerikanisch, Saddam, China, NATO	7%
	I-DE	SPD, Unternehmen, Euro, CDU, Kläger, Gewerkschaft, Partei, Kommune, Schröder, Grüne, Antrag, BGB, Bürger, PDS, Prozent, Regelung, kommunal, Landkreis, Beschäftigte,	11%
	НКРЯ	выбор, реформа, партия, государство, путин, регион, депутат, экономический, бюджет, чечня, ситуация, экономика, гражданин, система, считать, решение	18%
	I-RU	партия, государственный, международный, общество, народ, экономический, сша, выбор, путин, гражданин, общественный, суд, военный, орган, министр, ядерный, регион	10%
	I-RU	беларусь, белорусский, Лукашенко, республика, минск, президент, государство, предприятие, союзный, страна, экономический, республиканский, белоррусия, кгб	1%
	I-ZH	退党, **农**民, 党, 自由, 政府, 上**访**, 法制, **报导**, 江**泽**民, 人**权**, 胡**锦**涛, 警察, 毛**泽东**, 法律, 政**权**, 民众, **宪**法, 胡耀邦, 腐**败**, 中国人, **权**力, 公民, **记**者, **历**史, **权**利, **讯**, 宗教, **专**制, **镇压**, 革命	7%
	I-ZH	美元, 伊拉克, 表示, 官**员**, 国家, 日本, 政府, ,**韩**国, 人民**币**, **贸**易, 爆炸, **举**行, 俄, 台湾, **报导**, **战**争, **汇**率, **领导**人, 会**谈**, **联**合国, 分子, 国**际**, 俄**罗**斯, 地区, **进**行, **计**划, 阿富汗, **选举**, 伊朗 , 石油	6%
Comm	BNC	company, market, business, price, rate, cost, firm, share, tax, investment, turnover, prot, account, financial, goods, income, management, customer, interest	8%
	I-EN	Government, Act, UK, scheme, costs, services, income, Ireland, Wales, sector, MR, cent, financial, tax, local, per, Minister, ensure, public, goods, schemes, transport, policy, licence	20%
	I-DE	dpa, AG, Aktie, THW, Unternehmen, Bank, Dollar, Euro, Mark, GmbH, Million, Mio, Prozent, Berlin, Milliarde, Frankfurter, Hamburg, Stuttgart, Markt, Berliner, iwr, Quartal	7%
	НКРЯ	финансовый, акция, проект, директор, фирма, экономический, бюджет, средства, договор, нефть, продажа, экономика, акционер, стоимость, инвестор, продукция	3%
	I-RU	стоимость, фонд, рф, договор, финансовый, федерация, акция, сумма, проект, экономический, доход, инвестиция, товар, банк, нефть, имущество, суд, срок, бюджет	10%
	I-ZH	**银**行, 行**业**, **销**售, **开发**, 提升, 全国, **业务**, 投**资**, **设计**, 汽**车**, 需求, 品牌, 固定, 提供, **项**目, 期, 消**费**, 基金, 数据, **亿**, **标**准, 上市, 分析, 使用, 我国, **应**用, **电**子, 文化, 网友	10%
Comp	BNC	system, software, user, data, computer, ibm, unix, file, technology, program, database, version, package, disk, interface, microsoft, server, error, code, access, dec, operating, processing	11%
	I-EN	software,file, computer, digital, system, user, text, Windows, server, code, MPlayer, image, page, int, Internet, model, mobile, output, number, Fortran, BibTeX, interface, network,	9%
	I-DE	Datei, Software, Kunde, Rechner, Version, Internet, Server, Gerät , Windows, Benutzer, digital, Computer, Apple, verwenden, Linux, installieren, Anbieter, Datenbank, PC, speichern	10%
	I-RU	сеть, устройство, диск, сервер, функция, версия, сообщение, windows, компьютерный, программный, код, список, ноутбук, документ, Компьютерра, спам, компания, доступ	8%
Health	BNC	patient, cell, disease, gastric, molecule, treatment, species, gene, dna, protein, pylorus, bile, ulcer, mucosa, symptom, colitis, pollution, water, gall, serum, bowel, oesophageal	2%
	I-EN	blood, Health, hospital, treatment, care, disease, patient, baby, birth, mental, cells, cancer, women, medical, NHS, doctor, clinical, behaviour, HIV, pain, sexual, therapy, symptoms, risk	10%
	I-DE	Behandlung, Therapie, Medikament, Krankheit, Medizin, medizinisch, Klinik, Erkrankung, Gehirn, Symptom, Zelle, häfiug, Körper, Studie, Untersuchung, Laser, Diagnose, Störung	9%
	НКРЯ	секс, жена, акт, девушка, врач, девочка, партнер, интимный, ребенок, беременность, эрекция, брак, спид, мужской, женский, непридумывать, нравиться, пенис, сношение	2%
	I-RU	пациент, семья, мужчина, чувство, состояние, поведение, врач, сон, ситуация, частый, мать, процесс, испытывать, упражнение, детский, беременность, реакция, мышца	9%

Таблица 4. Сравнение тематической структуры на основе кластеризации.