Митрофанова О.А., Шиморина А.С.,
Кольцова О.Ю., Кольцов С.Н.

Моделирование семантических связей в текстах социальных сетей с помощью алгоритма LDA (на материале русскоязычного сегмента Живого Журнала)[1]

1. Введение

Создание сверхбольших корпусов текстов, разработка новых методов и алгоритмов лингвистического моделирования заставляет исследователей по-новому взглянуть на смысловую компрессию, специфический класс задач автоматической обработки естественного языка. Это задачи автоматического выделения ключевых слов и словосочетаний, классификации и кластеризации лексики и документов в корпусах текстов, выделения классов слов с близкими дистрибутивными свойствами и т.п. [Леонтьева 2006]. В данном ряду особняком стоит задача тематического моделирования корпусов текстов, поскольку пристальное внимание лингвистов и социологов сейчас обращено к анализу социальных сетей и выявлению тематической структуры сообществ [Bodrunova et al. 2013; Koltsova et al. 2011]. Компьютерная обработка корпусов текстов, сформированных на основе социальных сетей, открывает широкие возможности для оперативной оценки не только общественного мнения, но и состояния русскоязычного дискурса, динамики словаря, развития внутриязыковых связей.

Цель нашего исследования заключается в том, чтобы 1) осуществить эксперименты по моделированию тематики корпуса текстов Живого Журнала (ЖЖ) Livejournal.ru с помощью программного комплекса TopicMiner, основанного на алгоритме LDA (Latent Dirichlet Allocation), 2) определить содержательное наполнение тем, отраженных в записях пользователей ЖЖ, 3) выявить и проинтерпретировать основные типы семантических связей слов внутри тем, 4) найти адекватные лингвистические модели анализа полученных экспериментальных данных.

2. Моделирование тематики текстов в компьютерной лингвистике

Тематическое моделирование – способ построения модели корпуса текстов, отражающий переход от совокупности документов, совокупности слов в документах к набору тем, характеризующих содержание документов. Тематические модели – модели со скрытыми переменными, для выявления которых лучше всего подходит нечеткая кластеризация. При нечеткой кластеризации любое слово или документ с некоторой вероятностью относится к нескольким темам. Можно сказать, что в тематической модели

[1] Исследование выполнено совместными усилиями кафедры математической лингвистики СПбГУ и лаборатории ЛИНИС НИУ ВШЭ. В данной научной работе использованы результаты проекта «Социально-политические процессы в Интернете», выполненного в рамках Программы фундаментальных исследований НИУ ВШЭ в 2013 году.

текстовой коллекции описанию слова или документа ставится в соответствие семейство вероятностных распределений на множестве тем.

В практических разработках последних лет широко используется ряд методов тематического моделирования. Среди алгебраических моделей текста, на которые опираются процедуры тематического моделирования, наиболее распространены стандартная векторная модель текста VSM (Vector Space Model) и латентно-семантический анализ LSA (Latent Semantic Analysis), среди вероятностных (генеративных) моделей наиболее всего применяются вероятностный латентно-семантический анализ pLSA (probalilistic Latent Semantic Analysis), латентное размещение Дирихле LDA (Latent Dirichlet Allocation). Дадим их краткую характеристику (подробные обзоры см. [Anaya 2011; Blei, Ng, Jordan 2003; Daud et al. 2010; Lee, Song, Kim 2010; Воронцов, Потапенко 2012; Коршунов, Гомзин 2012], Topic Modelling Bibliography).

В VSM, а также и в других моделях, текст рассматривается как «мешок слов» и описывается терм-документной матрицей. Словам или текстам ставятся в соответствие вектора в n-мерном пространстве. Соответственно, сравнение векторов позволяет оценить, насколько слово характерно для некого документа, насколько связаны слова между собой в документе, насколько близки документы. Недостатки VSM связаны с неудобством использования в работе с текстами больших объемов, с невозможностью учесть синонимические отношения между словами и их многозначность. Модель LSA наследует основные характеристики VSM, однако в LSA для выявления наиболее значимых слов в текстах используется разложение терм-документной матрицы по сингулярным значениям – так называемое сингулярное разложение (Singular Value Decomposition, SVD). В рамках такого разложения можно выбрать K наибольших сингулярных значений, соответственно, в преобразованной терм-документной матрице останутся K первых линейно независимых компонент, что отражает основную структуру различных зависимостей, присутствующих в исходной матрице. Таким образом, каждый термин и документ представляются при помощи векторов в общем семантическом пространстве размерности K. Близость между любой комбинацией терминов и/или документов легко вычисляется при помощи скалярного произведения векторов. Значение величины K зависит от задачи: если K велико, то результаты метода близки к результатам VSM, если значение K мало, то данный метод не позволяет уловить различие между похожими документами или терминами. Недостатками метода LSA является сложность работы с большими разреженными матрицами, также данная методика позволяет указать лишь близость документов или терминов между собой, но не позволяет сгруппировать похожие документы – термины в темы.

Методы pLSA и LDA относятся к алгоритмам вероятностного тематического моделирования, которые позволяют анализировать слова в огромных наборах документов, а также выявлять скрытые темы, связи между темами и изменение их во времени. В основе pLSI лежит модель, которая связывает скрытые переменные тем с каждым наблюдаемым словом или документом. Таким образом, каждый документ может относиться к нескольким темам с некоторой вероятностью, что является отличительной особенностью этой модели по сравнению с подходами, не позволяющими вероятностного моделирования. Недостатки pLSA обуславливаются следующими причинами. Во-первых, данная модель содержит большое число параметров, которое растет в линейной зависимости от числа документов. Соответственно, модель склонна к переобучению и неприменима к большим наборам данных. В-вторых, отсутствует какая-либо закономерность при генерации документов из сочетания полученных тем. Все эти недостатки устранены в модели LDA, где порождение документа, характеризующих его тем, слов в этой теме производится с опорой на распределение Дирихле. Данные о свободно распространяемых пакетах для тематического моделирования на основе LDA приведены в конце статьи.

Известны также и другие тематические модели, в той или иной мере связанные с упомянутыми выше: это, в частности, совместная вероятностная модель JPM (Joint Probabilistic Model), скрытая тематическая марковская модель AHMM (Aspect Hidden Markov Model), автор-тематическая модель ATM (Author-Topic Model), модель автор-получатель ARTM (Author-Recipient Topic Model), корреляционная тематическая модель CTM (Correlated Topic Model) и т.п.

Спецификация тематических моделей может быть связана с различиями на уровне лингвистической обработки входных текстов и с привлечением аппарата дистрибутивной семантики [Baroni, Bernardi, Zamparelli, to appear; Mitchell, Lapata 2010]. Большинство разработчиков языковых ресурсов склоняются к необходимости интеграции статистических и традиционных лингвистических методов анализа, что было реализовано, например, в моделях семантического пространства WSM (Word Space Model), HAL (Hyperspace Analogue to Language), COALS (Correlated Occurrence Analogue to Lexical Semantics) и т.п. [Rohde, Gonnerman, Plaut 2005; Sahlgren 2006]. Использование многоуровневой разметки текстов (лемматизация, морфосинтаксическая и – в случае доступности – семантическая аннотация), учет границ синтаксических групп и существующих внутри них связей, ограничение контекстного окна, назначение весов контекстных элементов, нормализация значений коэффициентов совместной встречаемости, свертка признакового пространства – эти и другие методы способствуют повышению качества автоматической обработки текстов.

2. Корпусные данные и программное обеспечение экспериментов

Корпус текстов для проведения экспериментов по тематическому моделированию был автоматически сформирован на основе постов Живого Журнала Livejournal.ru. Корпус ЖЖ включает в себя записи первых 2000 блогеров по рейтингу популярности Живого Журнала за 4 недели (11.03.13 – 07.04.13), всего 103056 постов, около 30,5 млн с/у. Для загрузки текстов использовался компьютерный инструмент BlogMiner, созданный в ЛИНИС НИУ ВШЭ (разработчики О.Ю.Кольцова, С.Н.Кольцов).

Предобработка текстов корпуса ЖЖ включает в себя графематический анализ, лемматизацию, очистку от нетекстовых элементов (прежде всего, html-тегов), создание списка стоп-слов. Лемматизация текстов осуществлялась с помощью морфологического анализатора mystem [http://company.yandex.ru/technologies/mystem/, Segalovich 2003]. В предобработке использовался стоп-словарь объемом около 1300 лексем. В первую очередь к стоп-словам были отнесены закрытые классы слов — предлоги, союзы, междометия, частицы, местоименные форманты и вводные слова. Данные единицы более близки к грамматическим, нежели к лексическим средствам языка, и таким образом могут быть легко отброшены при тематическом моделировании. Основой для составления данного списка послужил словарь СССРЯ [СССРЯ 1997]. Дальнейшая обработка текстов показала, что в стоп-словарь нужно внести римские цифры, обозначаемые латинскими буквами, некоторые характерные для ЖЖ слова и обозначения (например, *ljuser*), а также нерусскоязычные (прежде всего, английские и украинские) слова. Сверх 1300 заранее выделенных лексем, в стоп-словарь были также добавлены слова, имеющие частоту менее 5 с/у в корпусе текстов. После очистки корпуса от стоп-слов его размер существенно уменьшается, остается около 40% от исходного объема.

В экспериментах задействован программный комплекс для тематического моделирования TopicMiner, разработанный в ЛИНИС НИУ ВШЭ (разработчики О.Ю.Кольцова, С.Н.Кольцов). TopicMiner позволяет проводить процедуры предобработки корпуса текстов и собственно тематического моделирования его содержания. Тематическое моделирование в TopicMiner проводится с помощью алгоритма латентного размещения Дирихле с сэмплированием Гиббса [Blei, Ng, Jordan 2003; Griffiths, Steyvers 2004]. Результатом работы программного комплекса являются списки наиболее вероятных документов и слов для каждой темы. Эксперименты по тематическому моделированию проводились в несколько итераций с изменением параметров. Число тем варьировалось от 50 до 400 с шагом 50 (50, 100, 150, 200, ... 400). Объем списков слов, соотносимых с темой, ограничивался по умолчанию 100 словами. Все слова внутри темы качественно равноправны. В силу того, что для лингвистической интерпретации результатов было

необходимо назначить метки тем, в качестве таковых выбирались слова, формально занимающие первую позицию в списке.

3. Лингвистическая интерпретация результатов экспериментов

Наибольший интерес для лингвистического анализа представляет содержательное наполнение тем, характеризующих корпус ЖЖ, а также исследование типов связей, которые эксплицированы в наборах слов, формирующих темы.

Слова, составляющие автоматически сформированные темы, распределяются между номинативным, атрибутивным и предикативным классами. Наиболее многочисленным является номинативный класс, который представлен существительными абстрактными и конкретными, нарицательными и собственными. Номинативный класс включает в себя обозначения общих понятий (*Бог, жизнь, мир, смерть* и т.п.), институтов и явлений общественно-политической жизни (*армия, власть, государство, закон, культура, музей, наука, общество, организация, партия, страна, фестиваль, церковь, экономика* и т.п.); человека (*девочка, дитя, друг, женщина, мама, муж, старик* и т.п.), животных (*животное, кот, кошка, олень, собака* и т.п.), бытовых реалий (*автомобиль, аэропорт, город, деревня, дом, здание, игрушка, квартира, корабль, магазин, масло, машина, одежда, отель, самолет, сахар, телефон, файл* и т.п.). Из имен собственных встречаются имена и фамилии (*Александр, Андрей, Джон, Иван* и т.п.; *Березовский, Навальный, Медведев, Проханов, Путин, Собчак, Сталин* и т.п.), топонимы (*Грузия, Италия, Кипр, Россия, США* и т.п.*; Екатеринбург, Москва, Петербург, Ростов* и т.п.). Атрибутивный класс включает качественные (*великий, древний, красивый, хороший* и т.п.) и относительные прилагательные (*английский, русский, японский* и т.п.). Предикативный класс оказывается самым узким, он представлен глаголами типа *выставлять, любить, оставлять, просить, прощать, рассказывать, смотреть, случаться, сообщить, считать* и т.п.

Автоматически сформированные темы различаются по информативности. С одной стороны, мы наблюдаем темы, в которых актуализируется основное содержание текстов корпуса ЖЖ (например, общественно-политическая проблематика в неформальном изложении, свойственном дискурсу социальных сетей), с другой стороны, это фоновые темы, которые формируются общей лексикой и присутствуют в любых текстах независимо от их тематики (например, *день, месяц, цвет* и т.п.). Если в социологических исследованиях более важны темы первого типа, отражающие общественное мнение в пределах определенного хронологического среза, то в лингвистическом аспекте полезен анализ того, какие семантические связи реализуются внутри тем обоих типов, а также оценка универсальности этих связей в русском языке.

Важный аспект в оценке содержания корпуса ЖЖ – это многовариантность тематического анализа, проявляющаяся в возможности пересечения тем. Например, при разных параметрах экспериментов повторяются темы с метками *власть, день, дорога, друг, жизнь, компания, мир, область, работа, Россия, самолет, слово, страна, фильм, хороший* и т.п. Это смежные темы, различающиеся наполнением и отражающие разные аспекты одного явления. Рассмотрим в качестве примера две темы с меткой *Россия*. Ниже в списках представлены первые 20 слов, упорядоченные по значению коэффициента ассоциации: *Россия1, Путин, партия, власть, выбор, депутат, президент, единый, Навальный, политическая, глава, оппозиция, политик, член, страна, Владимир, дума, митинг, кандидат, Медведев* и т.п.; *Россия2, страна, русская, народ, государство, власть, мир, общество, русский, советский, война, запад, право, история, национальный, СССР, российский, политик, западный, революция* и т.п. Можно предположить, что тема *Россия1* объединяет тексты о современной политической истории России, тогда как тема *Россия2* более связана с описанием советского периода.

Смежными являются темы с высокой долей общих слов. Максимальная доля совпадений составляет 36%, как например, в списках для тем *самолет* и *танк* (*тип, экипаж, управление, высокий, боевой, разработка, ракета, масса, система, комплекс, устанавливать, находиться, скорость, двигатель, технический, машин, вариант, тяжелый, установка, применение, разрабатывать, конструкция, техник, кг, создавать, км, частить, вооружение, боевая, дальность, составлять, работа, конструктор, оружие, испытание, целить*), что можно объяснить стереотипностью описаний боевой техники.

Анализируя связи слов внутри тем, мы принимали во внимание следующее соображение: состав тем определяется автоматически в результате построения статистической модели корпуса текстов, которая отражает близость дистрибутивных свойств формирующих тему слов, тенденцию их взаимозамены или совместного употребления. На этом основании отношения между словами в темах можно характеризовать как контекстные или квази-отношения, поскольку они наблюдаются в рамках определенного корпуса текстов.

Проиллюстрируем обработанный нами материал в табл. 1, где приведены слова из 20 случайно выбранных тем (списки ограничены первыми 20 позициями из 100).

Таблица 1. Слова из 20 случайно выбранных тем (первые 20 из 100 позиций)

бог¶	**жизнь¶**	**правда¶**	**церковь¶**	**ученый¶**
земля¶	мир¶	верить¶	православный¶	земля¶
душа¶	отношение¶	врать¶	папа¶	планета¶
мир¶	чувство¶	ложь¶	бог¶	космический¶
сердце¶	хороший¶	читать¶	храм¶	исследование¶
народ¶	друг¶	слово¶	святая¶	мир¶
дух¶	собственный¶	плохо¶	религиозный¶	звезда¶
сила¶	любовь¶	совесть¶	священник¶	источник¶
имя¶	сила¶	поверить¶	религия¶	космос¶
жизнь¶	желание¶	обманывать¶	христос¶	наука¶
великий¶	счастие¶	пора¶	святой¶	луна¶

любовь¶	жить¶	хороший¶	вера¶	находиться¶
господь¶	образ¶	лень¶	иисус¶	видеть¶
свет¶	страх¶	признаваться¶	отец¶	показывать¶
прощать¶	высокий¶	веселый¶	католический¶	время¶
рука¶	состояние¶	почитать¶	божий¶	научный¶
видеть¶	энергия¶	очередной¶	ватикан¶	теория¶
враг¶	время¶	понятнее¶	церковный¶	обнаруживать¶
господин¶	чувствовать¶	наглый¶	великий¶	энергия¶

закон¶	**власть¶**	**сталин¶**	**оружие¶**	**армия¶**
право¶	политическая¶	ленин¶	пистолет¶	военный¶
российский¶	страна¶	партия¶	винтовка¶	война¶
россия¶	государство¶	цк¶	патрон¶	солдат¶
рф¶	политик¶	товарищ¶	пуля¶	войска¶
суд¶	народ¶	рабочий¶	ствол¶	сила¶
организация¶	политический¶	хрущев¶	стрельба¶	генерал¶
документ¶	общество¶	кпсс¶	стрелять¶	полк¶
гражданин¶	интерес¶	брежнев¶	выстрел¶	бой¶
государственный¶	сила¶	власть¶	противник¶	офицер¶
федерация¶	борьба¶	советский¶	действие¶	командир¶
решение¶	национальный¶	член¶	пулемет¶	время¶
орган¶	лидер¶	время¶	ружье¶	противник¶
федеральный¶	движение¶	большевик¶	расстояние¶	служба¶
лицо¶	демократия¶	ссср¶	время¶	танк¶
принимать¶	элита¶	берия¶	легкий¶	фронт¶
деятельность¶	отношение¶	руководитель¶	магазин¶	действие¶
запрещать¶	государственный¶	андропов¶	приводить¶	бой¶
министр¶	демократический¶	секретарить¶	бой¶	операция¶
информация¶	партия¶	троцкий¶	первое¶	оборона¶

книга¶	**письмо¶**	**слово¶**	**писать¶**	**внимание¶**
слово¶	отправлять¶	язык¶	написать¶	обращать¶
язык¶	почта¶	русский¶	читать¶	привлекать¶
читать¶	заказ¶	фраза¶	прочитывать¶	следовать¶
автор¶	написать¶	буква¶	слово¶	замечать¶
текст¶	адрес¶	выражение¶	присылать¶	рука¶
русский¶	доставка¶	словарь¶	история¶	уделять¶
писать¶	посылка¶	мат¶	добрый¶	указывать¶
написать¶	присылать¶	речь¶	интересно¶	сторона¶
английский¶	сайт¶	перевод¶	письмо¶	палец¶
история¶	заказывать¶	звучать¶	жизнь¶	интересный¶
литература¶	почтовый¶	значение¶	доходить¶	проблема¶
писатель¶	просьба¶	английский¶	лекарство¶	обращаться¶
роман¶	сообщение¶	употреблять¶	интернет¶	стоить¶
перевод¶	просить¶	по¶	тема¶	ладонь¶
стих¶	доставлять¶	произносить¶	умирать¶	признак¶
книжка¶	почесть¶	термин¶	передавать¶	возникать¶
прочитывать¶	придти¶	смысл¶	сталин¶	отвлекать¶
читатель¶	писать¶	форма¶	выхаживать¶	система¶
выражение¶	электронный¶	русская¶	честно¶	состояние¶

работа¶	**газета¶**	**машина¶**	**самолет¶**	**производство¶**
художник¶	журналист¶	автомобиль¶	система¶	завод¶
картина¶	статья¶	дорога¶	проект¶	предприятие¶
искусство¶	интернет¶	водитель¶	корабль¶	производить¶
выставка¶	сми¶	место¶	машина¶	продукция¶
мастер¶	писать¶	ехать¶	двигатель¶	промышленный¶
портрет¶	новость¶	робот¶	вооружение¶	промышленность¶
создавать¶	пресса¶	номер¶	ракета¶	техник¶
работать¶	слово¶	ездить¶	установка¶	рабочий¶
рисунок¶	информация¶	стоить¶	производство¶	оборудование¶
рисовать¶	интервью¶	движение¶	мм¶	производитель¶
автор¶	канал¶	час¶	завод¶	фабрика¶
живопись¶	тема¶	хороший¶	техник¶	заказ¶
творчество¶	написать¶	колесо¶	танк¶	производственный¶
художественный¶	сайт¶	скорость¶	разработка¶	выпускать¶
мир¶	мк¶	время¶	оружие¶	цех¶
миллион¶	эфир¶	руль¶	время¶	технология¶
материал¶	мнение¶	задний¶	полет¶	рынок¶
продавать¶	издание¶	знак¶	работа¶	построить¶
произведение¶	публикация¶	остановка¶	вертолет¶	фирма¶

Соотношение слов в темах отражает многообразие парадигматических и синтагматических отношений, организующих текст. Нам представляется, что наиболее удобная схема описания языковых связей внутри тем – это аппарат лексических функций в модели «Смысл <=> Текст» [Мельчук 1974/1999; ТКС 1984], позволяющий охватить предсказуемые, идиоматизированные связи слова и его лексических коррелятов (парадигматических вариантов – «замен» и синтагматических партнеров – «параметров»).

Среди парадигматических отношений доминируют синонимия (Syn), антонимия (Anti), конверсия (Conv), гиперонимические (Gener), деривационные (Der) отношения и т.п. Например: Syn: *дитя – ребенок; церковь – храм; бог – господь; предприятие – производство; армия – войска; оружие – вооружение* и т.п.; Anti: *правда – ложь; хороший – плохой; детский – взрослый; брать – отдавать; зарабатывать – тратить (деньги); привлекать – отвлекать (внимание)* и т.п.; Conv: *обращать – привлекать (внимание)* и т.п.; Gener: *игра – футбол; оружие – пистолет, винтовка, пулемет, ружье; машина – автомобиль; животное – собака, кошка, кот, медведь* и т.п.; Der: *казак – казачий; Италия – итальянский; игра – играть, выигрывать; работать – работа; стрелять – стрельба* и т.п. Кроме того, внутри тем наблюдаются партитивные отношения: *семья – родители, ребенок, мама, отец, сын, дочь; армия – полк, генерал, офицер, солдат*; *винтовка – ствол, пуля, магазин; машина – руль, колесо; животное – хвост* и т.п.

Синтагматические отношения реализуются на уровне рамок валентностей, заполняемых словами из темы. Среди лексических функций этим связям соответствуют, например, функции Oper1,2, связывающие глагол, название первого или второго актантов в роли подлежащего и название ситуации в роли дополнения: *внимание – обращать, привлекать, уделять, отвлекать; письмо – отправлять, писать, написать, присылать, доставлять; фильм – снимать; дом – строить* и т.п. В именных атрибутивных сочетаниях, обозначающих характерный признак объекта, реализуется лексическая функция Ver («соответствующий назначению»): *власть – политическая, государственная; закон – российский, федеральный; животное – дикое; цвет – яркий* и т.п. Довольно распространены стандартные метонимические связи внутри тем, закрепляемые лексической функцией Attr: *водитель – машина, автомобиль; автор, писатель – книга; ученый – наука; журналист – газета; художник – картина* и т.п. Был обнаружен ряд примеров на реализацию лексической функции Cap («глава», «начальник»): *страна – президент; штаб – начальник; отряд – командир* и т.п., лексической функции Equip («личный состав»): *полк – солдат*; *стадо – олень* и т.п., лексической функции Doc(res) («документ, являющийся результатом»): *писать – письмо*; *рисовать – рисунок* и т.п.

Во многих случаях мы обнаружили, что наполнение тем соответствует традиционным лексико-семантическим или тематическим закрытым группам, например: *час – минута, секунда, сутки* и т.п.; *(месяц) – январь, февраль, март, апрель* и т.п., *знак (зодиака) – лев, весы, близнец, рак, рыба, дева, скорпион, телец, водолей* и т.п.

Полученный нами материал подтверждает, что внутренняя организация тем, описываемая как пучок парадигматических и синтагматических связей между словами из темы, проецируется на структуру отраженных в корпусе ЖЖ экстралингвистических ситуаций, где участникам и их признакам соответствует лексическое наполнение тем с учетом их сочетаемостных предпочтений, типового заполнения рамок валентностей и т.п. Поэтому внутренняя организация тем может быть представлена в виде ассоциативных сетей или ситуационных фреймов [Quillian 1968; Sowa 2000; Азарова 1989; Минский 1979; Филлмор 1988; Цейтин 1985], где в качестве имени ситуации будет выбрана метка темы, а представляющие тему слова будут распределены по слотам, соответствующим участникам ситуации и их признакам. Примеры ситуационных фреймов для тем *игра, производство, письмо* и *слово* приведены в табл. 2.

Таблица 2. Ситуационные фреймы для тем *игра, производство, письмо*, *слово*

имя фрейма: *игра* класс: *игра*... подклассы: *спорт, футбол,...* атрибуты: *спортивный…* действие: *играть, выигрывать…* действующее лицо: *игрок,...* организация: *команда, клуб…* событие: *чемпионат, матч…* результат: *победа…* …	**имя фрейма: *производство*** класс: *промышленность…* подклассы: *завод, предприятие, производство, фабрика, фирма…* компоненты: *цех…* действие: *производить, выпускать…* действующее лицо: *техник, рабочий…* инструмент: *оборудование…* результат: *продукция…* …
имя фрейма: *письмо* объекты: *письмо*, *посылка, сообщение…* атрибуты: *адрес, конверт…; почтовый, электронный…* действие: *писать, отправлять, получать, присылать…; заказ, доставка…* место: *почта…* действующее лицо: *клиент, курьер…* …	**имя фрейма: *слово*** классы: *язык, речь, словарь, …* подклассы: *фраза, выражение, слово,термин, буква…* атрибуты: *значение, смысл, форма…; английский, русский…* действие: *звучать, произносить, употреблять…; перевод…* …

В лингвистическом смысле тема организуется словами, проявляющими тенденцию к совместному употреблению в устойчивых языковых структурах, точнее говоря, в конструкциях с фиксированными лексическими компонентами [Fillmore 1986; Рахилина 2010]. Например, как конструкцию следует рассматривать сочетания типа V(*привлекать, обращать, уделять…*) + *внимание* + PR (*на, к, #…*) + ADJ(*интересный…*) + S(*проблема…*); V(*указывать…*) + PR (*на…*) + S(*проблема…*); S(*состояние…*) + S(*проблема…*) и т.п. Тем самым, тематическое моделирование дает ценный материал для исследований в области автоматического выделения конструкций [Lyashevskaya et al. 2011; Митрофанова и др. 2012].

На сегодняшний день в русскоязычной компьютерной лингвистике отсутствуют доступные программные средства, использующие LDA и сопоставимые с инструментом тематического моделирования типа TopicMiner. По этой причине для проверки данных,

полученных с помощью TopicMiner, мы сравнили наполнение тем со списками реакций на соответствующие стимулы в Русском ассоциативном словаре [РАС 2002]. Тексты, присутствующие в социальных сетях вообще и в ЖЖ в частности, есть результат языкового творчества разных авторов, это полилогический дискурс, созданный в условиях, которые приближаются к спонтанному ассоциативному эксперименту. Тем самым, можно предположить, что в корпусе ЖЖ содержатся ассоциативные связи, отражающие языковое сознание русскоязычных блогеров первой декады XXI века. Поскольку данные РАС относятся к 1980-м...1990-м годам XX века, сравнение наиболее продуктивных стимулов и связанных с ними ассоциаций, формирующих ядро языкового сознания русскоязычной среды конца XX века [Уфимцева 1998, 2002], с темами, автоматически сгенерированными при обработке корпуса ЖЖ, дает возможность выделить стабильные понятия, фигурирующие в дискурсе независимо от эпохи, и оценить динамику русскоязычной картины мира.

Так, в случае с темой *церковь* мы наблюдаем совпадение свыше 30% слов в составе темы и в списке ассоциатов из РАС (*православный, храм, священник, вера, святая, Христос, Иисус, бог, божий, икона, монастырь, христианский, религия, молитва, крест, епископ, приход, обряд* и т.п.) Это указывает на существование стабильной идиоматизированной области в парадигматических и синтагматических связях слова *церковь*, которые одновременно отражается и в теме (где преобладает парадигматика), и в данных ассоциативного эксперимента (с преобладанием синтагматики).

При сравнении данных корпуса ЖЖ и РАС для темы *газета* мы выявили не более 15% пересечений (*информация, интересно, ложь, новость, письмо, политический, правда, пресса, российская, статья, текст, хороший, читать* и т.п.). В составе темы *газета* присутствуют слова, связанные с электронными СМИ (*сайт, сеть, интернет, канал, эфир* и т.п.), тогда как в РАС отражено представление о газете как печатном периодическом издании (*свежая, вечерняя, новая, бумага, шрифт, киоск, ежедневная, еженедельная, запах краски, запах типографии, почтовый ящик* и т.п.; *Humanite, Morning Star, Аргументы и факты, Литературка, Коммунист, Киноафиша, Молодой учитель, Труд* и т.п.). Столь глубокое различие между корпусом ЖЖ и РАС объясняется не только методологическими расхождениями данных лингвистических источников, но и динамикой языкового сознания.

4. Заключение

Наблюдения, сделанные нами в ходе экспериментов, подтверждают целесообразность использования алгоритма LDA при решении задач тематического моделирования, в том числе при оценке содержания текстов в результате семантической компрессии. Данные, полученные при обработке корпуса текстов Живого Журнала с помощью компьютерного

инструмента тематического моделирования TopicMiner (ЛИНИС НИУ ВШЭ), свидетельствуют о многообразии тематики записей русскоязычных авторов ЖЖ. Нам удалось выявить содержательное ядро корпуса, описать его в виде набора тем, систематизировать семантические отношения между словами внутри тем, детально охарактеризовать парадигматические и синтагматические связи в темах. Исследовательский материал допускает интерпретацию с позиций теории лексических функций, ситуационной фреймовой семантики и грамматики конструкций. Содержательное наполнение тем позволяет делать выводы о динамике языкового сознания русскоязычных пользователей социальных сетей.

Перспективы исследования связаны с совершенствованием инструментов автоматической обработки текстов, с поиском методов автоматизации построения ситуационных моделей на основе тем, с проведением тематического моделирования в специализированных корпусах текстов.

Литература

Anaya L.A. Comparing Latent Dirichlet Allocation and Latent Semantic Analysis as Classifiers. University of North Texas, 2011. http://digital.library.unt.edu/ark:/67531/metadc103284/

Baroni M., Bernardi R., Zamparelli R. Frege in space: A program for compositional distributional semantics. Linguistic Issues in Language Technologies. [To appear]

Blei D.M., Ng A.Y., Jordan M.I. Latent Dirichlet Allocation // Journal of Machine Learning Research 3 (4–5), January 2003.

Bodrunova S., Koltsov S., Koltsova O., Nikolenko S.I., Shimorina A. Interval Semi-Supervised LDA: Classifying Needles in a Haystack // 12th Mexican International Conference, MICAI 2013, Mexico City, Mexico, November 24-30, 2013, Proceedings, Part I: Advances in Artificial Intelligence and Its Applications. LNCS, LNAI. Vol. 8265.

Daud A., Li J., Zhou L., Muhammad F. Knowledge Discovery through Directed Probabilistic Topic Models: a Survey // Proceedings of Frontiers of Computer Science in China. 2010.

Fillmore Ch.J. The Mechanisms of Construction Grammar // Proceedings of the Berkeley Linguistic Society. Vol. 14. 1988.

Griffiths T., Steyvers M. Finding Scientific Topics // Proceedings of the National Academy of Sciences. Vol. 101. 2004.

Koltsova O., Maslinsky K., Koltsov S. Protests, Elections and Their Contributions to the Topical Structure of the Russian Blogosphere: a «Big Data Approach» // Internet, Politics, Policy

2012: Big data, Big Challenges?, Oxford Internet University 20–21 September 2012. http://www.hse.ru/data/2012/12/19/1303704698/2.pdf

Lee S., Song J., Kim Y. An Empirical Comparison of Four Text Mining Methods // 43rd Hawaii International Conference on System Sciences HICSS 2010.

Lyashevskaya O., Mitrofanova O., Grachkova M., Romanov S., Shimorina A., and Shurygina A. Automatic Word Sense Disambiguation and Construction Identification Based on Corpus Multilevel Annotation // Text, Speech and Dialogue. Proceedings of the 14th International Conference TSD 2011, Pilsen, Czech Republic, September 1–5, 2011. Springer-Verlag, 2011.

Mitchell J., Lapata M. Composition in Distributional Models of Semantics. Cognitive Science, 34:8, 2010.

Quillian M.R. Semantic memory // Semantic Information Processing. MIT Press, Cambridge, Massachusetts, 1968.

Rohde D.L.T., Gonnerman L.M., Plaut D.C. An Improved Model of Semantic Similarity Based on Lexical Co-Occurrence. 2005. http://tedlab.mit.edu/~dr/Papers/RohdeGonnermanPlaut-COALS.pdf

Sahlgren M. The Word-Space Model: Using Distributional Analysis to Represent Syntagmatic and Paradigmatic Relations between Words in High-Dimensional Vector Spaces. Ph.D. dissertation, Department of Linguistics, Stockholm University. 2006. http://www.sics.se/~mange/TheWordSpaceModel.pdf

Segalovich I. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine // MLMTA–2003. http://download.yandex.ru/company/iseg-las-vegas.pdf

Sowa J.F. Knowledge Representation: Logical, Philosophical, and Computational Foundations. Brooks Cole Publishing Co., Pacific Grove, CA, 2000.

Topic Modelling Bibliography // http://www.cs.princeton.edu/~mimno/topics.html

Азарова И.В. Использование сетевых представлений лингвистических данных при автоматической обработке текста. Дис. ... канд. филол. наук. Л., 1989.

Воронцов К.В., Потапенко А.А. Регуляризация, робастность и разреженность вероятностных тематических моделей // Компьютерные исследования и моделирование. 2012. Т. 4. № 4.

Коршунов А., Гомзин А. Тематическое моделирование текстов на естественном языке // Труды ИСП РАН. М., 2012.

Леонтьева Н.Н. Автоматическое понимание текстов: Системы, модели, ресурсы. М., 2006.

Мельчук И.А. Опыт теории лингвистических моделей «Смысл <=> Текст». М., 1974 / 1999.

Минский М. Фреймы для представления знаний. М., 1979.

Митрофанова О.А., Ляшевская О.Н., Грачкова М.А., Шиморина А.С., Шурыгина А.С., Романов С.В. Эксперименты по автоматическому разрешению лексико-семантической неоднозначности и выделению конструкций (на материале Национального корпуса русского языка) // Структурная и прикладная лингвистика. Вып. 9. СПб., 2012.

РАС – Караулов Ю.Н., Черкасова Г.А., Уфимцева Н.В. и др. Русский ассоциативный словарь. Т.1–2. М., 2002.

Рахилина Е.В. (ред.) Лингвистика конструкций. М., 2010.

СССРЯ – Морковкин В.В. и др. Словарь структурных слов русского языка / Под ред. В.В. Морковкина. М., 1997.

ТКС – Мельчук И.А., Жолковский А.К. и др. Толково-комбинаторный словарь современного русского языка. Опыты семантико-синтаксического описания русской лексики. Вена, 1984.

Уфимцева Н.В. Этнический характер, образ себя и языковое сознание русских // Языковое сознание: формирование и функционирование. М., 1998.

Уфимцева Н.В. Ядро языкового сознания русских (по данным массовых ассоциативных экспериментов) // Корпусная лингвистика и лингвистические базы данных. СПб., 2002.

Филлмор Ч. Фреймы и семантика понимания // Новое в зарубежной лингвистике. Вып. XII. Когнитивные аспекты языка. М., 1988.

Цейтин Г.С. Программирование на ассоциативных сетях // ЭВМ в проектировании и производстве. Л., 1985.

Программные реализации LDA

http://code.google.com/p/topic-modeling-tool/

http://cran.r-project.org/web/packages/lda/index.html

http://cran.r-project.org/web/packages/topicmodels/index.html

http://mallet.cs.umass.edu/index.php

http://nlp.stanford.edu/software/tmt/tmt-0.4/

http://radimrehurek.com/gensim/intro.html

http://www.people.fas.harvard.edu/~ptoulis/code/LDA_Perl.zip

https://cwiki.apache.org/confluence/display/MAHOUT/Latent+Dirichlet+Allocation