Категоризация текстов для структурирования массива исторических документов

© Г.В. Артемова © К.К. Боярский © Н.Ф. Гусарова © Н.В. Добренко
Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики

© Е.А. Каневский
Санкт-Петербургский экономико-математический институт РАН
Санкт-Петербург

g.o.artemova@gmail.com boyarin9@yandex.ru natfed@list.ru
kanev@emi.nw.ru

Аннотация

Рассматриваются подходы к автоматическому выделению терминов узкой предметной области из текстов. На примере исторического кораблестроения показано, что в условиях ограниченной номенклатуры текстов их анализ стандартными средствами дает неудовлетворительные результаты. Для повышения качества выделения терминов предложено предварительно подвергать текст полному синтаксическому разбору с построением дерева зависимостей и переходить к дальнейшему анализу по укрупненным наборам лексем с близкой семантикой – темам. Показано, что при этом компьютерные оценки схожести тематики текстов приближаются к экспертным.

1 Введение

Эффективное использование информации, содержащейся в текстовых источниках, должно основываться на построении концептуальной модели изучаемой предметной области (ПрО). Наиболее универсальными и подходящими для совместного использования формами такого моделирования считаются онтологии. Известно определение, данное T.R. Gruber, согласно которому онтология является точной спецификацией концептуализации [3].

В работе решалась задача формирования корпуса текстов (категоризация текстов) для структурирования массива исторической информации с целью последующего автоматизированного построения онтологии на примере исторического кораблестроения.

Тексты по тематике исторического кораблестроения – сложный объект для обработки методами компьютерной лингвистики и машинного обучения. Они очень сильно различаются по длине (от 1–2 фраз до 1600 страниц полноформатного книжного текста), а также по частоте встречаемости специальной лексики. Например, слово «судно» (во всех словоформах) занимает четвертое место по частоте встречаемости, уступая только служебным словам «и», «на», «в» и опережая все слова общей лексики. В то же время имеется большой набор специальных терминов, важных для построения онтологии, частота встречаемости которых составляет 1–2 на фрагмент, т.е. находится на уровне лексического шума. Не выполняется гипотеза равномерного распределения лексем по тексту, важная для применения методов машинного обучения при обработке естественно-языковых текстов (NLP). Наконец, за 300 лет существования российского флота изменялись и сами корабли, и описывающая их лексика, что также нужно учитывать при построении онтологии.

В итоге традиционные методы категоризации текстов, основанные на ранжировании лексем по частоте встречаемости и сравнении соответствующих векторов для ПрО «Историческое кораблестроение» дают неудовлетворительные результаты. Эти неудачи вполне понятны. За десятилетия, разделяющие даты создания текстов по идентичной тематике, лексикон авторов сильно изменился. Кроме того, в силу специфики ПрО суммарный объем доступных для анализа текстов ограничен, что не позволяет применять традиционные для машинного обучения методы регуляризации, дающие прекрасные результаты на больших текстовых корпусах.

Таким образом, необходимо иметь формальный метод категоризации текстов по статистически

Труды 16-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2014, Дубна, Россия, 13–16 октября 2014 г.

однородным областям. Для этого в работе предложено перейти из векторного представления документов в пространстве лексем в векторное представление в пространстве тем (существенно уменьшив при этом размерность векторов). Эффективность этого приема подтверждена анализом литературных источников [1, 2, 4]. Однако специфика настоящей работы состоит в том, что в качестве базы для такого перехода использован семантический классификатор В.А. Тузова [9], который, в отличие от других классификаторов, может обеспечить однозначное соотнесение каждой лексемы с каким-либо классом.

2 Характеристики текстов

Для формирования понятийной структуры онтологии были отобраны тексты, отражающие структурирование предметной области, принятое в профессиональной среде в соответствующую эпоху. А именно, в качестве основных текстовых источников были использованы [7] и [8]. Для последующего применения методов NLP тексты подвергались предварительной обработке. По книге О. Курти [7] было выполнено сканирование и распознавание в программе Abbyy FineReader 11; последующего редактирования не потребовалось. По книге Ш. Ромма [8] автоматическое сканирование показало неудовлетворительные результаты, в связи с чем использован ручной набор с последующим редактированием. Текст приведен к современной орфографии, в том числе:

– устранен твердый знак в конце слов, заменены буквы ять, і и ижица;

– устаревшие глаголы и наречия заменены их современными синонимами, если таковые имеются и однозначны (например, «оканчевается» заменено на «оканчивается»).

Для исследования статистических свойств отобранных текстов из обеих книг были выделены близкие по тематике фрагменты (табл. 1), причем предпочтение отдавалось экспертным оценкам близости.

Таблица 1. Анализируемые фрагменты текстов

Обозначение фрагмента	Источник	Глава	Объем фрагмента (словоупотреблений)
[Ф1]	[7]	Глава 3 «Конструкция корпуса судна»	5400
[Ф2]	[7]	Глава 7 «Постройка корпуса модели судна»	8700
[Ф3]	[7]	Глава 10 «Паруса»	2400
[Ф4]	[8]	Глава 3 «О чертежах кораблей»	4100
[Ф5]	[8]	Глава 8 «О парусах и их оснащении»	5800

Как было установлено экспертными оценками, фрагменты содержательно образуют две тематические группы (категории): «Корпуса кораблей и их моделей» ([Ф1], [Ф2], [Ф4]) и «Паруса» ([Ф3], [Ф5]).

Для лемматизации и снятия омонимии в работе использован разработанный авторами парсер SemSin [6]. Парсер построен в рамках концепции dependency parsing и реализует лексикализованную модель разбора предложений. В качестве базы используется расширенный и модифицированный словарь В.А. Тузова [9]. В словарь добавлены термины предметной области, а в классификатор – три класса («Рангоут», «Парус», «Такелаж»), к которым отнесено свыше 200 слов. Для удобства восприятия текста классам были даны содержательные читаемые названия вместо числовых обозначений.

Работа парсера основана на применении набора продукционных правил [5], причем для каждого предложения парсер производит полный синтактико-семантический анализ и строит дерево зависимостей. Такой разбор текста позволил, во-первых, решить проблему омонимии и, во-вторых, уже на предварительном этапе исключить из рассмотрения не только стоп-слова, но и резко сузить круг возможных кандидатов в термины ПрО. На первом этапе работы в качестве таковых рассматривались только существительные, однако используемая технология позволяет легко изменять состав терминов, в том числе в виде групп слов, неконтактно стоящих в предложениях.

Результаты работы парсера SemSin по текстам исторического кораблестроения оказались вполне удовлетворительными. Так, точность определения лемм была не хуже 97% даже на исторических текстах. Например, без какой-либо предварительной настройки на предметную область система определяла, что в предложениях типа «На деревянных судах собственно водонепроницаемых переборок нет, но имеются выгородки» правильная лемма – «судно», а не «суд». Точность выявления существительных, которые в первую очередь рассматриваются как кандидаты в термины, была примерно такого же порядка. Заметим, что принятый подход упрощает дальнейшую работу по выявлению терминов, поскольку в любом тексте самыми частотными являются служебные слова – предлоги, союзы и т. д., а они в построенном словаре отсутствуют по определению.

3 Статистические показатели «по словам»

Эффективность предложенного метода категоризации текстов была проверена экспериментально. Для этого проведено сопоставление качества категоризации внутри каждой из выделенных тематических групп (табл. 1).

Категоризация проводилась двумя методами – традиционным (частотное выделение лексем) и предложенным (тематическое выделение лексем).

Таблица 2. Статистические показатели наиболее частотных слов фрагментов группы «Корпуса кораблей и их моделей»

		[Ф1]		[Ф2]		[Ф4]	
№	**Лексема**	**Частота**	**Вектор**	**Частота**	**Вектор**	**Частота**	**Вектор**
1	СУДНО	0.0635	0.6847				0
2	КОРПУС	0.0236	0.2546	0.0725	0.6859		0
3	ШПАНГОУТ	0.0257	0.2773	0.0428	0.4057	0.0368	0.3194
4	ОБШИВКА	0.0304	0.3282	0.0225	0.2131		0
5	МОДЕЛЬ		0.0000	0.0342	0.3240		0
6	ПАЛУБА	0.0241	0.2603	0.0213	0.2014		0
7	ДОСКА	0.0188	0.2037	0.0240	0.2277		0
8	ЧАСТЬ	0.0230	0.2490				0
9	КИЛЬ	0.0183	0.1981	0.0154	0.1459		0
10	ТОЛЩИНА		0.0000	0.0209	0.1985		0
11	ПЛАНКА		0.0000	0.0191	0.1810		0
12	БИМС	0.0167	0.1811		0.0000		0
13	КОРМА	0.0131	0.1415		0.0000		0
14	СЕЧЕНИЕ			0.0178	0.1693		0
15	ПЛОСКОСТЬ					0.0475	0.4117
16	РЫБИНА			0.0169	0.1605	0.0458	0.3975
17	ДЕРЕВО			0.0157	0.1489	0.0393	0.3407
18	ТОЧКА					0.0376	0.3265
19	ОБВОД			0.0151	0.1430	0.0327	0.2839
20	ФИГУРА			0.0135	0.1284	0.0327	0.2839
21	ЛИНИЯ			0.0120	0.1138	0.0319	0.2768
22	ЛЕКАЛО			0.0114	0.1080	0.0278	0.2413
23	ПРОЕКЦИЯ			0.0114	0.1080	0.0262	0.2272

Были выделены наиболее частотные слова во фрагментах группы «Корпуса» (табл. 2), где приведены частоты их встречаемости в процентах и построены нормированные векторы, компоненты которых d_i определялись по формуле

$$d_i = \frac{p(w_i)}{\sqrt{\sum_i p(w_i)^2}},$$

где p_i – частота появления i-го слова в данном фрагменте.

Проверена гипотеза о том, что частотный состав словаря определяет меру тематической близости текстов. В качестве интегральной оценки сходства текстов принята косинусная мера:

$$sim\left(\mathbf{d}_i, \mathbf{d}_j\right) = \cos\left(\angle\left(\mathbf{d}_i, \mathbf{d}_j\right)\right) = \frac{\sum_k d_{ik} d_{jk}}{\sqrt{\sum_k d_{ik}^2} \sqrt{\sum_k d_{jk}^2}}.$$

На основании данных табл. 2 для фрагментов [Ф1] и [Ф2] получено значение cos φ = 0.48, а для фрагментов [Ф1] и [Ф4] cos φ = 0.09, т. е. тематическое сходство фрагментов практически отсутствует, что, безусловно, не совпадает с экспертной оценкой.

Такой же анализ выполнен для фрагментов группы «Паруса». В табл. 3 отобраны наиболее частотные лексемы, покрывающие 45% всех словоупотреблений. Для [Ф3] это 26 лексем с частотностью семь и выше, для [Ф5] – 25 лексем с частотностью 14 и выше. Оказалось, что из отобранных лексем только 7 встречаются в обеих выборках. По отобранным лексемам (табл. 3) построены векторы и рассчитана тематическая близость фрагментов [Ф3], [Ф5]. Получено значение cosφ = 0,63, а если исключить лемму «парус», то cosφ = 0,25. Иначе говоря, расчет демонстрирует некоторое отдаленное сходство между выбранными фрагментами, но оно базируется фактически на одном слове. По остальной лексике связь очень низкая.

На примере фрагмента [Ф5] оценена частота вхождения специальных терминов (названий парусов) в анализируемые тексты. Оказалось, что среди 15 названий парусов только один (грот) попадает в описанную выше выборку и еще три имеют частотность выше трех (лисель, грот-марсель и марсель). Многие специальные термины имеют малую частоту (до 1–2 на фрагмент).

4 Статистические показатели «по темам»

Представленные выше результаты позволяют сформулировать проблемы, возникающие при кластеризации текстов рассматриваемой ПрО:

- Даже среди лексем с большим весом есть те, которые вряд ли могут стать терминами (вода, фут, пересечение).
- Некоторые специальные термины имеют малую частоту (до 1–2 на фрагмент). В то же время среди малочастотных лексем есть как явные кандидаты на термины предметной области (фок-мачта, катер, парусник), так и случайные слова

(варвар, время), что еще больше затрудняет отбор. В результате кандидаты в термины останутся незаметными среди лексического шума.

- Не выполняется условие равномерного распределения лексем по тексту.

Для преодоления этих проблем предлагается в задаче классификации перейти из векторного представления документов в пространстве слов в векторное представление в пространстве тем, уменьшив тем самым размерность векторов. Но тогда возникает новая проблема – как соотнести слово и тему.

Таблица 3. Статистические показатели наиболее частотных слов фрагментов группы «Паруса»

[Ф3]				[Ф5]			
Лексема	**Вхождений**	**Частота**	**Вектор**	**Лексема**	**Вхождений**	**Частота**	**Вектор**
ПАРУС	116	0.2843	0.8722	ПАРУС	111	0.1438	0.5850
СУДНО	27	0.0662	0.2030	БЛОК	70	0.0907	0.3689
РЕЙ	24	0.0588	0.1804	ФИГ	50	0.0648	0.2635
КЛИВЕР	22	0.0539	0.1654	ВЕРЕВКА	48	0.0622	0.2530
ШКАТОРИНА	20	0.0490	0.1504	УГОЛ	48	0.0622	0.2530
УГОЛ	17	0.0417	0.1278	КОНЕЦ	46	0.0596	0.2424
ПАРУСИНА	15	0.0368	0.1128	ЛИК-ТРОС	41	0.0531	0.2161
БИЗАНЬ	14	0.0343	0.1053	ЛЮВЕРС	37	0.0479	0.1950
СТОРОНА	12	0.0294	0.0902	ЛИК	36	0.0466	0.1897
МАЧТА	11	0.0270	0.0827	ДЛИНА	24	0.0311	0.1265
ПАРУСА	10	0.0245	0.0752	ШКАТОРИНА	23	0.0298	0.1212
СТАКСЕЛЬ	10	0.0245	0.0752	ГРОТ	20	0.0259	0.1054
КРАЙ	9	0.0221	0.0677	ЧАСТЬ	20	0.0259	0.1054
ФОРМА	9	0.0221	0.0677	ПОЛОСА	19	0.0246	0.1001
БУЛИНЬ	8	0.0196	0.0601	ПОЛОТНИЩЕ	19	0.0246	0.1001
ДЕТАЛЬ	8	0.0196	0.0601	СНАСТЬ	19	0.0246	0.1001
КОНЕЦ	8	0.0196	0.0601	ШКОТ	19	0.0246	0.1001
ЛЕЕР	8	0.0196	0.0601	КОРАБЛЬ	17	0.0220	0.0896
ЛИКТРОС	8	0.0196	0.0601	ОБРАЗ	16	0.0207	0.0843
МАРСЕЛЬ	8	0.0196	0.0601	ПЛАНКА	16	0.0207	0.0843
РИФ	8	0.0196	0.0601	ВЫСОТА	15	0.0194	0.0790
РИФ-ГАТ	8	0.0196	0.0601	ГАЛС	15	0.0194	0.0790
ГРОТ	7	0.0172	0.0526	ПРЯДЬ	15	0.0194	0.0790
ЛЮВЕРС	7	0.0172	0.0526	НОК	14	0.0181	0.0738
РИФ-СЕЗЕНЬ	7	0.0172	0.0526	СЕРЕДИНА	14	0.0181	0.0738
ШПРЮЙТ	7	0.0172	0.0526				

Таблица 4. Результаты тематической группировки лексем в текстах группы «Корпуса кораблей и их моделей»

	[Ф1]		[Ф2]		[Ф4]	
Классы	**Слов**	**Вектор**	**Слов**	**Вектор**	**Слов**	**Вектор**
Плавсредства (судно, корвет)	140	0.2610	42	0.0789	26	0.1340
Части_судов (бушприт, киль, корма)	498	0.9286	402	0.7556	102	0.5255
Помещения (каюта, крюйт-камера)	73	0.1361	4	0.0075	1	0.0052
Фрагмент_помещения (настил, переборка)	44	0.0820	26	0.0488	12	0.0618
Детали (заклепка, кронштейн)	53	0.0988	188	0.3533	4	0.0206
Люки (люк, отверстие)	45	0.0839	51	0.0958	0	0.0000
Стройматериалы (балясина, брусок)	84	0.1566	162	0.3045	69	0.3555
Чертежн_инструменты (карандаш, лекало)	4	0.0075	16	0.0300	46	0.2370
Границы (край, кромка)	12	0.0224	17	0.0319	50	0.2576
Фигуры (плоскость, полукольцо)	17	0.0317	138	0.2593	118	0.6080
Обвод (обвод)	1	0.0019	11	0.0206	40	0.2061
Дерево (дерево, дуб)	13	0.0242	16	0.0300	39	0.2009
Корпус (корпус, основа)	16	0.0298	150	0.2819	2	0.0103
Модель (модель)		0.0000	111	0.2086	2	0.0103

Решить эту задачу, причем в полностью автоматическом режиме, позволяет использование семантических классов классификатора, так как в этом случае одновременно с разбором предложения и построением дерева зависимостей производится соотнесение каждой лексемы с каким-либо классом.

В табл. 4 приведены результаты тематической группировки лексем в текстах группы «Корпуса».

В табл. 5 приведены оценки сходства между фрагментами в группе. Рассчитывалась косинусная мера сходства (4), однако сравниваемые векторы d_i получены для нижней части таблицы путем частотного выделения лексем (табл. 2), а для верхней части – путем тематического выделения лексем (табл. 4).

Таблица 5. Оценка сходства между фрагментами

По темам / По словам	[Ф1]	[Ф2]	[Ф4]
[Ф1]	–	0.57	0.87
[Ф2]	0.48	–	0.76
[Ф4]	0.1	0.40	–

Табл. 5 показывает, что при тематическом выделении лексем мера сходства существенно выше, чем при частотном выделении, т.е. тематическое выделение лексем более соответствует экспертному мнению, чем частотное выделение.

Аналогичная проверка выполнена на группе текстовых фрагментов с тематикой «Паруса». Результаты тематического выделения лексем в текстах группы «Паруса» и соответствующие им векторы показаны в табл. 6.

Таблица 6. Результаты тематического выделения лексем в текстах группы «паруса»

	[Ф3]		[Ф5]	
Классы	**Слов**	**Вектор**	**Слов**	**Вектор**
Паруса	262	0.8375	175	0.4414
Части судов	161	0.5146	322	0.8122
Положение			48	0.1211
Плавсредства	38	0.1215		0.0000
Снасти	28	0.0895	104	0.2623
Фигуры	26	0.0831	68	0.1715
Сторона			47	0.1186
Блоки			54	0.1362
Ткани	20	0.0639		0.0000

По результатам табл. 6 рассчитана тематическая близость фрагментов [Ф3] и [Ф5]. Получено значение $\cos\varphi = 0{,}83$, а если исключить лемму «парус», то $\cos\varphi = 0{,}81$. Сравнивая эти значения с аналогичными данными, полученными выше на основании табл. 3, можно видеть, что использование тематического выделения лексем существенно повышает устойчивость результатов и характеризует совокупную семантическую близость текстов, а не только их совпадение по одному слову.

5 Заключение

Таким образом, как показали эксперименты, во всех случаях предложенный метод давал результаты, гораздо более близкие к экспертному мнению, чем традиционный метод.

Применение предложенного метода категоризации текстов путем тематического выделения лексем для построения онтологии ПрО обеспечивает ряд преимуществ:

– выявляются тексты, характеризуемые совокупной семантической близостью, а не просто совпадением по отдельным, хотя и очень частотным, терминам;

– существенно повышается устойчивость категоризации; что позволяет легко добавлять к уже сформированным категориям новые текстовые фрагменты, расширяя тем самым базу для построения онтологии;

– обработанные фрагменты содержат не только почти всех кандидатов на термины, но и очень мало шума в виде посторонних слов.

Не представляет принципиальных сложностей расширение списка терминов за счет двух- и трехсловных сочетаний, полученных из дерева зависимостей. Это существенно облегчает задачу дальнейшего построения онтологии

Литература

[1] Jamal A. Nasir, Iraklis Varlamis, Asim Karim, George Tsatsaronis. Semantic smoothing for text clustering // Knowledge-Based Systems, 54 (2013), 216–229.

[2] Mary-Claire van Leunen and Richard Lipton. How to have your abstract rejected. Newman D. et al. Automatic evaluation of topic coherence // Human Language.

[3] Neches et al. 1991] Neches et al. Enabling Technology for Knowledge Syaring. AI Magazine, Winter, 1991, 35–56.

[4] Yanjun Li, Soon M. Chung, John D. Holt. Text document clustering based on frequent word meaning sequences Data & Knowledge Engineering, 64 (2008), 381–404.

[5] Боярский К.К., Каневский Е.А. Язык правил для построения синтаксического дерева // Интернет и современное общество: Материалы XIV Всероссийской объединенной конференции «Интернет и современное общество». – СПб.: ООО «МультиПроджектСистемСервис», 2011. – С. 233–237.

[6] Каневский Е.А., Боярский К.К. Семантико-синтаксический анализатор SemSin // Международная конференция по компьютерной лингвистике «Диалог-2012», Бекасово, 30 мая – 3 июня 2012 г. http://www.dialog-21.ru/digest/2012/?type=doc

[7] Курти Орацио. Постройка моделей судов. Энциклопедия судомоделизма / сокр. пер. с итал. А.А. Чебана. – Л.: Судостроение, 1977. – 544 с.

[8] Ромм Шарль. Морское искусство или Главные начала и правила, научающие искусству строения, вооружения, правления и вождения кораблей. Часть 1 / пер. с франц. А.А. Шишков. – Типография Морского шляхетского кадетского корпуса. – Часть 1, 1793. 542 с. Часть 2, 1795. 355 с.

[9] Тузов В.А. Компьютерная семантика русского языка. – СПб.: Изд-во С.-Петерб. ун-та, 2004. – 400 с.

Text Categorization for Generation of Historical Shipbuilding Ontology

G. Artemova, K. Boyarsky, N. Gusarova, N. Dobrenko, E. Kanevsky

Approaches to automatic term extraction in the narrow subject domain from texts are considered. On the example of historical shipbuilding it is shown that in the conditions of the limited nomenclature of texts their analysis using standard means yields unsatisfactory results. For improvement of quality of term extraction we offer to fulfil full syntactic analysis of the text starting with creating of dependences tree and then passing to further analysis of integrated sets of lexemes (subjects) having close semantics. It is shown that thus computer estimates of similarity of subject of texts come nearer to the expert estimates.