А.С. Михайлов, Т.В. Соколова, А.А. Чеповский, А.М. Чеповский

Выявление тематической направленности текстов на естественных языках

Аннотация. В статье представлено описание разработанной методики тематической рубрикации текстов, основанной на использовании специализированных словарей. Методика применяется для текстов СМИ и коротких комментариев сети Интернет. Показана эффективность использования псевдооснов слов, выделенных аналитическими алгоритмами анализа словоформ, для задачи тематической рубрикации текстов на различных естественных языках и применимость предложенной методики на практике.

Ключевые слова: автоматическая рубрикация текстов, специализированный словарь, выделение псевдооснов.

Введение

Задача тематической, жанровой классификации и рубрикации является актуальной для создания систем автоматической обработки текстов на естественных языках [1-4]. Интерес представляют как алгоритмы решения данной задачи, так и выбор тех дифференцирующих признаков, которые определяют отнесение текстов к заданным рубрикам [1, 2, 5].

Одной из таких задач является, например, задача атрибуции текстов. По результатам проведенных в [6] экспериментов констатировано, что для задачи определения авторства текстов использование пар подряд идущих в тексте букв дает более точные результаты, чем использование слов и словосочетаний [2, 6]. Выдвинуто предположение о том, что в двухбуквенных сочетаниях частично отображаются структуры словоупотреблений в тексте – различного типа морфемы. Следующим шагом [2] было предложено использовать трехбуквенные сочетания в качестве стилевых признаков при решении задачи атрибуции текстов.

Обобщением ситуации является показанный принцип применимости частотных характеристик буквосочетаний для задач определения языков и текстов [5, 8]:

- частотные характеристики букв, двухбуквенных и трехбуквенных сочетаний являются дифференцирующими признаками для задач классификации текстов по языкам текстов и кодировкам;
- частотные характеристики буквосочетаний, начиная с четырехбуквенных, являются дифференцирующими признаками для задач тематической, стилевой и жанровой классификации текстов.

Использование результатов анализа синтаксических структур, авторских инвариантов (проценты содержания служебных слов) применимо только для больших, как правило, литературных текстов [2]. Методы нейронных сетей, метод опорных векторов требуют большие обучающие выборки, что ограничивает работу информационных систем в режиме реального времени при решении задач тематической направленности коротких текстов [2, 3].

Актуальной проблемой является классификация по тематической, психолингвистической направленности коротких текстов, состоящих из одного – двух предложений, или даже нескольких слов. Такие тексты встречаются в комментариях к интернет-блогам, на форумах. Такие задачи практически не решаются в условиях автоматической обработки текстов.

Таким образом, наиболее актуальной и трудной задачей является выбор дифференцирующих признаков для задач классификации и рубрикации текстов по тематической и психо-

лингвистической направленности текстов на различных языках [1, 7]. Выбор дифференцирующих признаков является ключевой проблемой для создания методик выявления нарушений в текстах на естественных языках сети Интернет. Отметим, что задача классификации и рубрикации текстов чаще всего привязана к конкретному естественному языку [2-4, 6, 8]. Поэтому представляет интерес определение таких универсальных методов и дифференцирующих признаков, которые могут рассматриваться с позиций единых методик для различных естественных языков.

Исследования различных грамматических форм в качестве дифференцирующих признаков для тематической классификации текстов на русском языке проводились в [1, 5, 8].

Данная работа посвящена разработке методики тематической рубрикации текстов для различных естественных языков с целью выявления нарушений в сети Интернет и выбору дифференцирующих признаков для данной задачи.

Будем рассматривать задачу рубрицирования текстов, понимая под этим отнесение текстов на естественном языке к одной или нескольким предопределенным категориям (рубрикам) по результатам анализа текстов на основе специализированных тематических словарей.

1. Специализированные словари

Создание специализированных тематических словарей происходит в два этапа. На первом этапе вручную лингвистами, специалистами по обработке информации создаются словники, содержащие основные слова и словосочетания по заданным тематикам. На втором этапе формируются специализированные тематические словари с использованием компьютерных морфологических словарей и программ морфологического анализа. Словари содержат данные о парадигме каждого слова.

Запись в специализированных тематических словарях – фраза (phrase), под которой понимается слово или словосочетание, входящее в одну или несколько рубрик. Слова представлены в словаре всей своей парадигмой и псевдоосновами слов, которые формируются автоматически при составлении специализированного словаря соответствующими морфологическими словарями и аналитическими алгоритмами морфологического анализа для различных естественных языков [9, 10]. Вхождению фразы в рубрику приписан числовой уровень, означающий степень соответствия фразы рубрике. Поддерживаются целочисленные значения уровня от 1 до 3: 1 означает слабое соответствие, 3 – сильное соответствие, 2 – промежуточное. В Табл. 1 приведены основные критерии, в соответствии с которыми проводилось разбиение лексики в словарях по уровням.

В тематических словарях содержится информация о языке и уровне ключевых слов и словосочетаний. Указания языка используется для ограничения поиска слов и словосочетаний только в текстах на соответствующем языке. Список составленных тематических словарей для четырех языков и их размеры приведены в Табл. 2. Под размером понимается общее количество фраз. Тематики определялись из задачи определения нарушений законодательства в текстах в мировой сети электронной коммуникации и носят несколько условные названия.

Наличие одного, двух и трех уровней в специализированных словарях определяется во многом лингвистической типологией языков.

Табл.1. Уровни слов и словосочетаний в словаре

3 уровень	Слова, устойчивые фразы и словосочетания, которые наиболее часто встречаются в текстах данной тематики, напрямую связаны с данной тематикой и характеризуют ее.
2 уровень	Слова и фразы данной тематики, но встречаются одновременно и в текстах на другие близкие тематики. Слова и фразы довольно часто встречаются в текстах данной тематики, характеризуют ее, но напрямую отнести их к тематике семантически нельзя. А также слова общей лексики – психические и эмоционально окрашенные клише, которые часто встречаются в данной теме и воспринимаются именно как относящиеся к данной теме.
1 уровень	Слова, фразы, клише, мемы, имена собственные, которые можно отнести к общеупотребительной лексике, но которые часто встречаются в текстах данной тематики и , как правило, носят негативный характер.

Табл. 2. Специализированные тематические словари

	Название тематического словаря	Размер словаря, записей		
		1 уровень	2 уровень	3 уровень
	русский язык			
1.	Экстремизм	1249	766	310
2.	Терроризм	1060	254	322
3.	Национализм, социальная рознь	207	205	311
4.	Фашизм	92	239	187
5.	Отрицание традиционных ценностей	188	49	287
6.	Насилие, жестокость	214	460	261
7.	Наркотики	1293	4074	643
8.	Порнография	77	346	291
	английский язык			
1.	Экстремизм	–	241	569
2.	Терроризм	–	795	465
3.	Национализм	–	302	675
4.	Нацизм, фашизм	–	809	523
5.	Нарушение прав, свобод	–	811	439
6.	Отрицание традиционных ценностей	–	592	379
	башкирский язык			
1.	Экстремизм	–	–	1067
2.	Терроризм	–	–	2078
3.	Национализм	–	–	873
4.	Сепаратизм	–	–	285
	татарский язык			
1.	Экстремизм	–	–	543
2.	Терроризм	–	–	613
3.	Национализм	–	–	302
4.	Сепаратизм	–	–	309

Русский язык является языком флективного типа, для которого характерна многозначность грамматических морфем, составляющих словоформы. Такое словообразовательное многообразие приводит к наличию различных словосочетаний (как именных, так и глагольных групп), несущих разнообразную смысловую нагрузку. Английский язык является языком аналитического типа, что сокращает количество лексических значений для конкретных словоформ, словосочетания имеют более четкие значения и отпадает необходимость в словаре первого уровня. Башкирский и татарский языки являются языками агглютинативными, для которых характерна грамматическая однозначность словообразовательных и словоизменительных морфем. Поэтому не возникает необходимости в учете словосочетаний с различными смысловыми сочетаниями, и для таких языков достаточно словаря третьего уровня.

На формирование словарей для таких языков как башкирский и татарский оказывают влияние ограничения связанные с небольшим количеством ресурсов в сети Интернет на этих языках, что определяет недостаточное количество текстов, слабую «разработанность» языка и приводит к малому количеству лексики, описывающей проявления нарушений законодательства, слабому разнообразию употребляемых выражений.

При анализе текста на первом этапе происходит разбиение текста на отдельные лексемы и определяется язык отдельных слов текста [7, 11]. Далее осуществляется поиск словоформы в тематическом словаре. Если словоформа из текста совпала с какой-либо словоформой из словаря, то считается, что в тексте встретилось соответствующее слово. Словарь содержит псевдоосновы, выделенные аналитическими алгоритмами морфологического анализа в автоматическом режиме [9, 10]. Это дает возможность заменять поиск по словоформе поиском по псевдоосновам, полученным для всех словоформ, выделенных в тексте в результате лек-

сического анализа с применением тех же аналитических алгоритмов морфологического анализа, которые используются при формировании словаря.

Таким образом, после лингвистического анализа формируется список найденных в тексте словосочетаний с указанием их позиций в тексте, рубрик, к которым они относятся, и уровней соответствия. В результате для каждой рубрики получается список непересекающихся фраз, псевдооснов слов и словосочетаний.

2. Оценка принадлежности текста рубрике

Определение рубрики осуществляется вычислением оценки принадлежности текста рубрике на основе анализа выделенных в текстах объектов (слов или словосочетаний), содержащихся в специализированных тематических словарях. Выделение объекта в тексте осуществляется либо по словоформам, принадлежащим данной фразе словаря, либо по псевдоосновам, отнесенным к конкретной фразе словаря.

Важно заметить, что далеко не всегда весь текст посвящен одной тематике, так же как и нарушение представления информации может содержаться лишь в небольшой его части. Для выявления таких участков оценку принадлежности текста рубрике будем вычислять не для целого текста, а для «окон» – содержащих текст интервалов $\boldsymbol{\Omega}$ фиксированной длины с заданным началом и концом. Оценка принадлежности текста рубрике определяется для «окна», начало которого совпадает с началом исследуемого текста, проводится оценка принадлежности текста в нем каждой из рубрик. Затем начало «окна» сдвигается на некоторую фиксированную величину и процедура оценки повторяется.

Каждому объекту в «окне» ставится в соответствие его вес. Если выделенный объект не встречался ранее в данном окне, то в зависимости от уровня словаря в качестве его веса берется значение $\boldsymbol{w_1^D, w_2^D}$ или $\boldsymbol{w_3^D}$ (для 1-ого, 2-ого и 3-его уровня соответственно).

Чем выше уровень словаря, тем больший вес имеют объекты (слова, словосочетание, псевдоосновы), присутствующие в нем:

$$\boldsymbol{w_1^D < w_2^D < w_3^D} \qquad (1)$$

Вес слова из словаря 1-го уровня зависит от общего количества выделенных в окне объектов 2-го и 3-го уровней. Это объясняется тем, что в этом словаре содержится преимущественно та лексика, которую можно отнести к общеупотребительной, но которая в то же время часто встречается в текстах данной тематики. Поэтому при наличии в тексте объектов из словаря более высокого уровня (2-го и 3-го) объекты 1-го уровня должны иметь больший вес:

$$\boldsymbol{w_1^D} = \begin{cases} \mathbf{1} + (\boldsymbol{m} - \mathbf{1})\left(\frac{N_3 + N_2}{wsize - N_1}\right)^{\frac{1}{s}}, & \boldsymbol{N_1 \neq wsize} \\ \mathbf{1}, & \boldsymbol{N_1 = wsize} \end{cases} \qquad (2)$$

где $\boldsymbol{N_k}$ - количество объектов $\boldsymbol{k}$ -ого уровня в окне, $\boldsymbol{wsize}$ - размер окна,

$\boldsymbol{w_1^D} \in [\mathbf{1}, \boldsymbol{m}]$, $\boldsymbol{m < w_2^D}$.

Чем выше доля объектов 2-ого и 3-его уровня в окне, тем ближе $\boldsymbol{w_1^D}$ к $\boldsymbol{m}$.

Величины $\boldsymbol{w_2^D}$, $\boldsymbol{w_3^D}$, $\boldsymbol{m}$ и $\boldsymbol{s}$ выступают в качестве параметров, их значения определяются по результатам тестирования для каждой из рубрик.

Иногда слово или словосочетание присутствует в окне два или более раз. В этом случае объекту следует назначать меньший вес, чем аналогичным объектам, присутствующим в окне на предыдущих позициях. Будем считать, что вероятность отнесения текста к рубрике выше, если в тексте присутствуют два различных слова из тематического словаря, чем если в тексте было выделено только одно слово из того же словаря, но оно встретилось дважды. Поэтому введем эмпирическую оценку уменьшения веса w_i^o несколько раз встречающегося объекта:

$$\boldsymbol{w_i^o = \left(\sqrt{b_i + 1} - \sqrt{b_i}\right) w_{d_i}^D} \qquad (3)$$

где $\boldsymbol{b_i}$ - количество раз, которое $\boldsymbol{i}$-тый объект уже встретился ранее в исследуемом окне,

$\boldsymbol{d_i}$- уровень словаря, к которому принадлежит $\boldsymbol{i}$-тый объект (принимает значение от 1 до 3).

Формула (3) обеспечивает монотонное убывание веса объекта таким образом, что если некоторый объект из словаря k -ого уровня присутствует в окне n раз, то суммарный вес всех его вхождений в окно будет равен $\sqrt{n}\, w_k^D$.

Итоговая оценка $\boldsymbol{f}$ принадлежности текста в окне $\boldsymbol{\Omega}$ к рубрике вычисляется как нормированная сумма весов всех находящихся в исследуемом окне объектов, содержащихся в соответствующем тематическом словаре:

$$\boldsymbol{f} = \frac{\sum_{i=1}^{N} \boldsymbol{w_i^o}}{\boldsymbol{w_3^D}\, \boldsymbol{wsize}},\ \boldsymbol{N = N_1 + N_2 + N_3},\ \boldsymbol{f} \in [\boldsymbol{0}, \boldsymbol{1}] \quad (4)$$

Формулы (3) и (1) задают значение числителя (4) меньше значения $\boldsymbol{Nw_3^D}$, а $\boldsymbol{N \le wsize}$, что обеспечивает принадлежность значения $\boldsymbol{f}$ отрезку [0, 1].

Текст будем считать относящимся к рубрике, если максимальная из оценок $\boldsymbol{f}$ по всем пройденным окнам превышает установленный для данной рубрики порог $\boldsymbol{p}$. Таким образом, формулируем критерий:

$$\boldsymbol{\max_j f_j > p} \quad (5)$$

Каждый полученный текст проверяется на соответствие с каждой рубрикой словаря независимо. При этом текст проверяется только на соответствие тем рубрикам, которые относятся к данному языку.

Проверка принадлежности текста данной рубрике $\boldsymbol{h}$ производится путем разбиения текста на окна $\boldsymbol{\Omega}$ размером $\boldsymbol{wsize}$ *с*лов со сдвигом $\boldsymbol{displacement}$ слов (если сдвиг меньше размера окна, то окна получаются перекрывающимися). Таким образом в качестве окон берутся интервалы:

$$[\boldsymbol{0}, \boldsymbol{wsize}), [\boldsymbol{displacement}, \boldsymbol{displacement} + \boldsymbol{wsize}), [\boldsymbol{2\ displacement}, \boldsymbol{2\ displacement} + \boldsymbol{wsize}), \ldots. \quad (6)$$

Размеры окна выбираются: $\boldsymbol{wsize = 40}$ и сдвига $\boldsymbol{displacement = 10}$.

Выделение объектов производится с использованием словарей сразу для всего текста. При этом для каждого выделенного объекта сохраняется его позиция в тексте. Обработка каждого окна на соответствие некоторой рубрике производится независимо. Все объекты данной рубрики не принадлежащие окну отфильтровываются (берутся все объекты данной рубрики, попавшие в окно). После этого для данного окна независимо рассчитывается предложенная формула с параметрами $\boldsymbol{m, w_2, w_3, s}$ соответствующими рассматриваемой рубрике. Получаем значение оценки $\boldsymbol{f_{\Omega r}}$ характеризующее степень соответствия окна и данной рубрики. В качестве показателя соответствия текста рубрике $\boldsymbol{r}$ берется максимум значений $\boldsymbol{f_{\Omega r}}$ по всем окнам данного текста.

Если для хотя бы одного окна в тексте значение $\boldsymbol{f_{\Omega r}}$ превышает пороговое значение $\boldsymbol{p}$ для данной рубрики $\boldsymbol{r}$, то текст считается соответствующим рубрике $\boldsymbol{r}$. Для практического использования методики определяются рекомендованные параметры для каждой рубрики, показавшие лучший результат на контрольной выборке по результатам экспериментов.

3. Показатели качества рубрикации

Для оценки качества рубрикации подсчитываются количества отнесенных к рубрике текстов: $\boldsymbol{r^+}$, $\boldsymbol{r^-}$, $\boldsymbol{o^+}$, $\boldsymbol{n^+}$, где буква отражает решение эксперта, знак результат автоматической рубрикации ($\boldsymbol{r}$ - рубрика, для которой проводится оценка качества; $\boldsymbol{o}$ - прочие рубрики; $\boldsymbol{n}$ - нейтральный текст, «+» означает принадлежность рубрике, «−» означает не принадлежность рубрике).

Таким образом:

$\boldsymbol{r^+}$ **-** количество текстов, которые были отнесены экспертом к рубрике $\boldsymbol{r}$ и были отнесены к ней по результатам автоматической рубрикации.

$\boldsymbol{r^-}$ **-** количество текстов, которые были отнесены экспертом к рубрике $\boldsymbol{r}$, но не были отнесены к ней по результатам автоматической рубрикации.

$\boldsymbol{o^+}$ **-** количество текстов, которые были отнесены экспертом к одной из рубрик $\boldsymbol{o}$**,** но по результатам автоматической рубрикации были отнесены к рубрике $\boldsymbol{r}$**.**

$\boldsymbol{n^+}$ **-** количество текстов, которые были отнесены экспертом к нейтральным текстам $\boldsymbol{n}$, но по результатам автоматической рубрикации были отнесены к рубрике $\boldsymbol{r}$.

Качество рубрикации определяется характеристиками: полнота, точность и усредненная точность.

Полнота $\boldsymbol{R}$ показывает, какая доля текстов, отнесенных экспертом к данной рубрике, была правильно рубрицирована.

$$\boldsymbol{R} = \frac{\boldsymbol{r^+}}{\boldsymbol{r^+ + r^-}} \quad (7)$$

Точность $\boldsymbol{P_0}$ показывает, какая доля текстов, отнесенных к $\boldsymbol{r}$ при автоматической рубрикации, действительно относится к данной рубрике, согласно мнению эксперта без учета текстов, отнесенных экспертами к другой рубрике:

$$\boldsymbol{P_0} = \frac{\boldsymbol{r^+}}{\boldsymbol{r^+ + n^+}} \quad (8)$$

Точность $\boldsymbol{P_1}$ показывает, какая доля текстов, отнесенных к $\boldsymbol{r}$ при автоматической рубрикации, действительно относится к данной рубрике, согласно мнению эксперта:

$$P_1 = \frac{r^+}{r^+ + o^+ + n^+} \quad (9)$$

Точность P_2 показывает, какая доля текстов, отнесенных к r при автоматической рубрикации, была отнесена экспертом к одной из рубрик:

$$P_2 = \frac{r^+ + o^+}{r^+ + o^+ + n^+} \quad (10)$$

Для интегральной оценки будем использовать усредненную точность $F_{\beta,\propto}$, вычисляемую как гармоническая средняя точности P_α и полноты R:

$$F_{\beta,\propto} = \frac{(\beta^2 + 1)P_\propto R}{\beta^2 P_\propto + R} \quad (11)$$

Усредненная точность (11) обычно называют F- мерой.

Точность P_1 (9) является для данной задачи точностью в классическом ее понимании, а точность P_2 (10) вводится дополнительно для более полной оценки качества. Такой шаг обусловлен особенностями решаемой задачи и позволяет лучше определять характер ошибок системы рубрикации. Из формул (9) и (10) видно, что вычисление точности P_2 проводится по аналогии с вычислением точности P_1, однако в первом случае отнесение любого текста к рубрике r, не отнесенного к ней экспертами, признается ошибкой, а во втором ошибкой считается только отнесение системой рубрикации нейтрального текста к рубрике r.

Близость некоторых из рассматриваемых рубрик и лексики, используемой в текстах на соответствующую им тематику, затрудняет получение значений точности P_1 близких к единице. Но предполагается, что текст, прошедший через систему рубрикации, в случае отнесения к одной из рубрик с нарушениями передается для повторной проверки и принятия итогового решения эксперту. Таким образом, например, система рубрикации со средними показателями точности P_1, но с высокими показателями точности P_2 может быть применима на практике и будет показывать отличные результаты при использовании группой экспертов.

4. Определение параметров

Выбор параметров осуществляется на обучающем наборе текстов, уже прошедший через модули предобработки. Таким образом, обрабатываются не тексты целиком, а их представления - список выделенных в тексте объектов с указанием их позиции в тексте, уровней и тематики словарей, в которых они содержатся.

Процедура выбора параметров состоит из следующих шагов:

1. Генерация всех возможных комбинаций параметров в задаваемых пользователем пределах и с указанным шагом с учетом наложенного ограничения:

$$m < w_2^D < w_3^D. \quad (12)$$

2. Вычисление значений по формуле (4) для всех наборов параметров, полученных на предыдущем шаге, для каждого текста из тестовой выборки.

3. Подбор наилучшего порога p для каждого набора параметров, т.е. такого значения p, при котором усредненная точность $F_{\beta\propto}$, достигает максимума.

4. Выбор наилучшей пары «набор параметров + порог». Лучшей считается пара с наибольшим значением $F_{\beta\propto}$,. Если таких пар несколько, выбирается любая.

При выборе параметров использовались $wsize = 40$ (размер окна) и $displacement = 10$ (сдвиг), а значения параметров варьируются от 1 до 30 для m, w_2^D,w_3^D и от 1 до 3 для s с шагом 1.

Для определения оптимальных параметров оценки принадлежности текста рубрике для текстов на русском языке был создан контрольный набор из текстов, предварительно отнесенных экспертами к одной из рубрик. Всего было использовано 310 текстов, из которых 100 текстов рассматриваются как нейтральные (не отнесенные ни к одной из рубрик), а остальные отнесены к одной из рубрик. Преобладание нейтральных текстов в наборе объясняется спецификой решаемой задачи: в Интернете доля текстов без нарушений заметно превышает долю таковых с нарушениями. Все тексты в наборе имеют размер более 30 слов.

В Табл. 3 и Табл. 4 приведены результаты выбора параметров для рубрикации текстов на русском языке с использованием словоформ (Табл. 3) и с использованием псевдооснов (Табл. 4). В таблицах указаны рекомендованные параметры для каждой рубрики, показавшие лучший результат на контрольной выборке, и оценка усредненной точности (11) данными параметрами на контрольной выборке. При вычислении усредненной точности

Табл.3. Рекомендованные параметры: словоформы

	m	w_2^D	w_3^D	s	p	$F_{0.5,1}$
Наркотики	3	4	29	1	0,025	0,921
Насилие, жестокость	9	23	27	2	0,05	0,316
Национализм, социальная рознь	17	19	23	3	0,06	0,729
Отрицание традиционных ценностей	2	8	28	2	0,025	0,793
Порнография	9	20	22	2	0,025	0,986
Терроризм	9	22	23	2	0,025	0,943
Фашизм	9	23	27	2	0,025	0,942
Экстремизм	9	25	29	1	0,05	0,855

Табл. 4. Рекомендованные параметры: псевдоосновы

	m	w_2^D	w_3^D	s	p	$F_{0.5,1}$
Наркотики	3	5	28	3	0,045	0,902
Насилие, жестокость	9	23	28	1	0,055	0,814
Национализм, социальная рознь	9	16	27	1	0,07	0,782
Отрицание традиционных ценностей	17	22	23	3	0,055	0,776
Порнография	9	23	28	2	0,04	0,982
Терроризм	10	14	15	3	0,04	0,88
Фашизм	9	24	26	2	0,035	0,948
Экстремизм	4	16	26	3	0,1	0,804

предпочтение отдавалось точности: в (11) использовался параметр $\boldsymbol{\beta} = \frac{1}{2}$.

Из Табл. 3 и Табл. 4 видно, что на стадии подбора параметров на обучающей выборке для псевдооснов и для словоформ удается достичь приблизительно одинаковых значений показателя усредненной точности $\boldsymbol{F_{\beta\propto}}$. В некоторых случаях для псевдооснов (например, рубрика «Насилие, жестокость») получаем более эффективную процедуру рубрикации, чем для варианта словоформ. Таким образом, на данном этапе наглядно показано, что для решении задачи рубрикации возможно использовать псевдоосновы.

5. Результаты экспериментов

На вход системе рубрикации подается текст, для которого на выходе получаем список рубрик, к которым этот текст был отнесен, с указанием ключевых фрагментов текста. Затем текст передается экспертам для дальнейшей оценки и принятия окончательного решения.

Тестирование полученной системы рубрикации текстов проводилось на тестовом наборе текстов на русском языке. Каждая рубрика была представлена в наборе в среднем 25 текстами («чистый» объем текстов – 960Kb). Результаты рубрикации текстов из тестового набора представлены для случая использования в качестве дифференцирующих параметров словоформ в Табл. 5, а для случая использования в качестве дифференцирующих параметров псевдооснов в Табл. 6.

В Табл. 7 и Табл. 8 приведены результаты подбора параметров и рубрицирования на тестовом наборе для татарского языка. В Табл. 7 представлены результаты для использования словоформ в качестве дифференцирующих параметров, а в Табл. 8 в качестве дифференцирующих параметров выбраны псевдоосновы, полученные алгоритмом автоматического выделения псевдооснов, описанным в [10]. Усредненная точность (11) подсчитана для точности $\boldsymbol{P_0}$. Точность $\boldsymbol{P_0}$ используется по причине ограниченного характера текстов на анализируемом языке и сложности создания тестирующих выборок, что повлияло и на низкое значение характеристики полноты при экспериментах.

Из анализа таблиц (Табл. 5-Табл. 8) следует, что даже при низкой точности $\boldsymbol{P_1}$ для конкретной рубрики достигается высокая точность $\boldsymbol{P_2}$ отнесения текста к потенциально содержащем нарушения. Таким образом, достигается задача выявления нарушений в автоматическом режиме для последующего анализа текста экспертами.

Табл. 5. Результаты рубрикации на тестовом наборе по словоформам

Рубрика	Точность P_1	Точность P_2	Полнота	$F_{1,1}$
Наркотики	0,75	0,938	0,8	0,774
Насилие, жестокость	0,668	1	0,567	0,613
Национализм, социальная рознь	0,546	1	0,6	0,572
Отрицание традиционных ценностей	0,684	1	0,433	0,531
Порнография	0,778	1	0,933	0,848
Терроризм	0,421	1	0,8	0,552
Фашизм	0,652	1	0,75	0,698
Экстремизм	0,538	0,97	0,65	0,588

Табл. 6. Результаты рубрикации на тестовом наборе по псевдоосновам

Рубрика	Точность P_1	Точность P_2	Полнота	$F_{1,1}$
Наркотики	0,743	0,857	0,867	0,8
Насилие, жестокость	0,69	1	0,533	0,601
Национализм, социальная рознь	0,771	1	0,9	0,831
Отрицание традиционных ценностей	0,634	0,957	0,431	0,513
Порнография	0,651	0,953	0,933	0,767
Терроризм	0,682	1	0,79	0,732
Фашизм	0,786	1	0,55	0,647
Экстремизм	0,722	1	0,521	0,605

Табл. 7. Результаты подбора параметров и рубрикации по словоформам для татарского языка

Рубрика	w_3^D	p	Точность P_0	Точность P_2	Полнота	$F_{1,0}$
Национализм	1	0,055	0,846	0,946	0,55	0,667
Сепаратизм	1	0,025	0,667	0,898	0,5	0,572
Терроризм	1	0,025	0,813	0,932	0,65	0,772
Экстремизм	1	0,045	0,394	0,71	0,65	0,491

Табл. 8. Результаты подбора параметров и рубрикации по псевдоосновам для татарского языка

Рубрика	w_3^D	p	Точность P_0	Точность P_2	Полнота	$F_{1,0}$
Национализм	1	0,08	0,857	0,941	0,6	0,705
Сепаратизм	1	0,06	0,452	0,764	0,7	0,784
Терроризм	1	0,095	0,9	0,963	0,45	0,600
Экстремизм	1	0,095	0,262	0,556	0,85	0,401

Из сравнения Табл. 5 с Табл. 6 и Табл. 7 с Табл. 8 следует, возможность использовать псевдоосновы в качестве дифференцирующих признаков, что является не менее эффективным, чем использование всех словоформ парадигмы слова из специализированного словаря.

Экспериментальные результаты показывают удовлетворительную точность для определения тематических рубрик, которые демонстрируют не столько фактографический, сколько психолингвистический характер текстов, их социологический характер.

Заключение

Разработана методика, на основе которой может быть построена система рубрикации текстов, позволяющая упростить и в то же время повысить эффективность работы экспертов по выявлению возможных нарушений в сети Интернет. Основная идея методики заключается в использовании специализированных тематических словарей, лексика в которых разбита по нескольким уровням.

Главный экспериментальный результат заключается в том, что данная методика рубрицирования позволяет с очень высокой точностью (P_2) определять возможные нарушения в публикациях средств массовой информации.

Задача рубрикации решалась для четырех языков: русский, английский, башкирский и татарский. Но описанная методика применима и для текстов на других языках при наличии соответствующих словарей.

Анализ результатов тестирования показывает, что эффективнее использовать псевдоосновы, выделенные аналитическим алгоритмом морфологического анализа для записей в тематических словарях, для автоматического рубрицирования коротких текстовых сообщений по специальным тематикам, связанным с нарушениями предоставления информации в Интернете. Таким образом, составляющие псевдоосновы морфемы содержат наиболее важную для эффективного решения задачи тематической рубрикации информацию о семантических свойствах лексем естественного языка, их определяющих психологическую направленность текста свойств. Использование псевдооснов позволяет говорить об общей для различных естественных языков методике рубрицирования, не требующей словарных морфологий для составления парадигм слов.

Литература

1. Бабенко М., Куршев Е., Одинцов О., Сулейманова Е., Чеповский А. Система классификации текстов информационных сообщений на русском языке "АКТИС" // В кн.: Труды международной конференции "Программные системы: теория и приложения", ИПС РАН, г. Переславль-Залесский, май 2004. — М.: Физматлит, 2004. Т.2. С.7-20.
2. Батура Т.В. Формальные методы определения авторства текста. Вестник НГУ. Серия: Информационные технологии. 2012. Том10. Выпуск 4. С. 81-94.
3. Мбайкоджи Э., Драль А.А., Соченков И.В.. Метод автоматической классификации коротких текстовых сообщений. // Информационные технологии и вычислительные системы. – 2012. – №3. С.93-102.
4. Боярский К.К., Каневский Е.А., Саганенко Г.И. К вопросу автоматической классификации текстов. Экономико-математические исследования: математические модели и информационные технологии.VII - СПб: СПб ЭМИ РАН. Нестор-История. 2009 2009. С. 252-273.
5. Гусев С.В., Поляков И.В., Чеповский А.М. Применение статистической модели текста в информационных системах // В кн.: Ершовская конференция по информатике 2011. Рабочий семинар «Наукоемкое программное обеспечение». — Новосибирск: Институт систем информатики им. А.П. Ершова, 2011. С. 69-72.
6. Кукушкина О.В., Поликарпов А.А., Хмелев Д.В. Определение авторства текста с использованием буквенной и грамматической информации // Проблемы передачи информации. М.: Наука, 2001. Т. 37, № 2. С. 96-108.
7. Чеповский А.М. Вопросы обработки текстовых сообщений на естественных языках // В кн.: SCVRT2013-14 Труды Международной научной конференции Международного центра по ядерной безопасности Института физико-технической информатики. Протвино: Изд-во ИФТИ, 2014. С. 250-254.
8. Андреев А.М., Березкин Д.В., Сюзев В.В., Шабанов В.И. Модели и методы автоматической классификации текстовых документов. — Вестник МГТУ им. Н.Э. Баумана. Сер. «Приборостроение», 2003, № 4. С. 64-94.
9. Болховитянов А.В., Чеповский А.М. Методы автоматического анализа словоформ // Информационные технологии. 2011. № 4 (176). С. 24-29.
10. Болховитянов А. В., Чеповский А. М. Алгоритмы морфологического анализа компьютерной лингвистики. Учебное пособие. — М.: МГУП им. Ивана Федорова, 2013. — 198 с.
11. Гусев С.В., Чеповский А.М. Модель для идентификации естественного языка текста // Бизнес-информатика, 2011. № 3(17). С. 31-35

Михайлов Александр Сергеевич. Руководитель службы ФГУП "ГлавНИВЦ" Управления делами Президента Российской Федерации. Окончил Московский институт стали и сплавов в 1998 году. Область научных интересов:автоматизированные системы управления и обработки информации.

Соколова Татьяна Владимировна. Выпускница бакалавриата Национального исследовательского университета "Высшая школа экономики".Область научных интересов: компьютерная лингвистика, обработка и анализ информации

Чеповский Александр Андреевич. Начальник отдела Национального исследовательского университета "Высшая школа экономики". Окончил Московский государственный университет им. М.В. Ломоносова в 2008 году. Кандидат физико-математических наук, доцент. Автор 8 печатных работ. Область научных интересов: компьютерные науки, теория колец, биоинформатика. E-mail: aachepovsky@hse.ru

Чеповский Андрей Михайлович. Доцент кафедры Информационной безопасности Национального исследовательского университета "Высшая школа экономики. Кандидат технических наук (1990). Автор свыше 80 печатных работ. Область научных интересов: кибернетика, компьютерная лингвистика, хранилища данных больших объемов.
E-mail: achepovskiy@hse.ru