УДК 004.021
DOI: 10.15827/0236-235X.120.678-683
Дата подачи статьи: 21.03.17
2017. Т. 30. № 4. С. 678–683

СПОСОБЫ ПРЕДСТАВЛЕНИЯ ТЕКСТОВОЙ ИНФОРМАЦИИ ПРИ АВТОМАТИЗИРОВАННОМ РУБРИЦИРОВАНИИ КОРОТКИХ ТЕКСТОВЫХ ДОКУМЕНТОВ

П.Ю. Козлов, аспирант, originaldod@gmail.com
(Смоленский филиал Национального исследовательского университета МЭИ, Энергетический проезд, 1, г. Смоленск, 214013, Россия)

Электронные сообщения граждан (жалобы, обращения, предложения и т.д.) с точки зрения возможности их автоматизированной обработки обладают рядом специфических особенностей: в значительной части случаев небольшой объем документа, что затрудняет его статистический анализ; отсутствие структуризации, что усложняет процедуры извлечения информации; наличие большого количества грамматических и синтаксических ошибок, что обусловливает необходимость реализации нескольких дополнительных этапов обработки; нестационарность тезауруса (состава и важности слов), зависящего от выхода новых нормативных документов, выступлений должностных лиц и политических деятелей и т.д., что вызывает необходимость использования процедур динамической классификации рубрик.

В статье описываются этапы автоматизированного анализа и методы формализации текстовых документов. Предлагается метод рубрицирования, который использует результаты морфологического и синтаксического этапов с модифицированной лингвистической разметкой текстовых документов.

В качестве синтаксического парсера рассматриваются современные программные продукты MaltParser и LinkGrammar, которые строят деревья зависимостей для всех предложений в документе. Приводятся стандартные лингвистические разметки MaltParser и LinkGrammar применительно к коротким текстовым документам, а также модификация разметки LinkGrammar для использования их рубрицирования.

В процессе использования известных программных продуктов для проведения дополнительных этапов анализа придется столкнуться с проблемой разнообразия лингвистических разметок. Например, большинство синтаксических парсеров на выходе представляет каждое предложение текста в виде деревьев зависимостей, которые описывают лингвистической разметкой. Лингвистическую разметку для дальнейшей классификации и назначения весовых коэффициентов необходимо модифицировать, тем самым увеличивая размерность метрики.

Описывается разработанный метод рубрицирования, который учитывает экспертную оценку важности слов для каждой рубрики, а также синтаксическую роль слов в предложениях. Приведена диаграмма процесса автоматизированного рубрицирования жалоб и предложений в разработанной системе анализа. Описан эксперимент, который подтверждает целесообразность использования синтаксических парсеров в подобных системах, что приводит к увеличению точности рубрицирования.

Даны рекомендации по улучшению точности разработанного метода и использованию аппарата теории нечетких множеств и методов когнитивного моделирования для разрешения проблемы нестационарности тезауруса систем, которые зависят от выхода нормативных документов и выступлений должностных лиц.

Ключевые слова: *автоматизированный анализ текстов, динамичный тезаурус, методы формализации текстовых документов.*

Одним из основных направлений государственной политики в Российской Федерации является повышение степени открытости органов государственной и муниципальных властей различных уровней, в том числе на основе организации их виртуального взаимодействия с населением. В результате происходит процесс постоянного совершенствования интернет-порталов органов исполнительной и законодательной власти, где каждый гражданин может в электронном виде подать жалобу, обращение или предложение. Число подобных электронных контактов непрерывно растет. Например, в Смоленской области за 2016 год поступило 9 936 электронных обращений, что составляет 30 % от всего числа жалоб [1]. За 2015 год в администрацию Санкт-Петербурга поступило 134 387 документов (из них 38 000 в электронном виде), за 2016 год – 117 274 документа (из них 54 473 в электронном виде) [2].

С учетом жестко регламентированных сроков направления ответа возникает необходимость обеспечения автоматизированной обработки указанных запросов с целью их рубрицирования для повышения оперативности взаимодействия с профильными структурными подразделениями администраций.

Электронные сообщения граждан (жалобы, обращения, предложения и т.д.) с точки зрения возможности их автоматизированной обработки имеют ряд специфических особенностей:

– небольшой объем большинства документов, что затрудняет его статистический анализ;

– отсутствие структуризации, что усложняет процедуры извлечения информации;

– наличие большого количества грамматических и синтаксических ошибок, обусловливающее необходимость реализации нескольких дополнительных этапов обработки;

– нестационарность тезауруса (состава и важности слов), который зависит от выхода новых нормативных документов, выступлений должностных лиц и политических деятелей и т.д., приводящая к необходимости использования процедур динамической классификации рубрик.

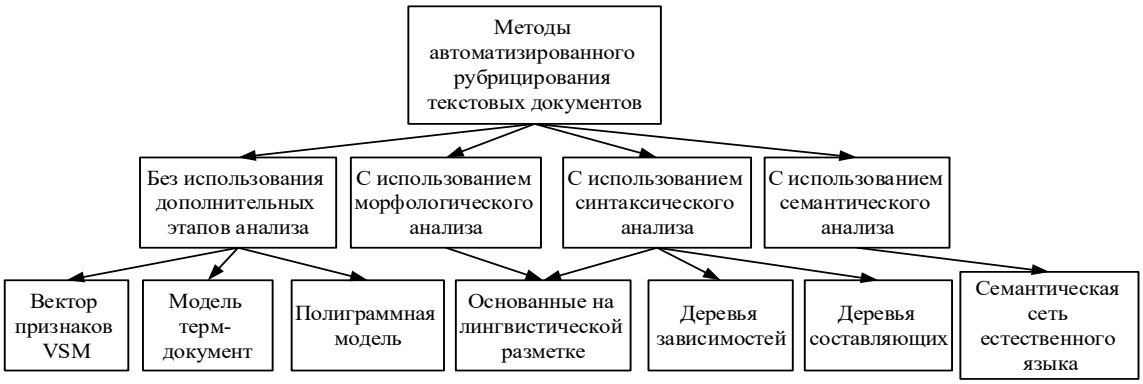

Рис. 1. Методы формализации текстовых документов на различных этапах анализа

Fig. 1. Methods for formalizing text documents at various stages of analysis

Очевидно, что указанные особенности рассматриваемых текстовых документов накладывают определенные ограничения на алгоритмы применения морфологического, синтаксического и семантического анализов, а также на соответствующие им процедуры формализации для автоматизированной обработки текстов, в том числе в рамках виртуальных систем информационного обеспечения различных региональных социально-экономических процессов (см, например, [3, 4]).

На рисунке 1 показаны основные известные способы формализации текстовых документов для последующей автоматизированной обработки с использованием различных алгоритмов.

В методах, где не используются дополнительные этапы анализа, достаточно представить текстовый документ в виде моделей VSM, терм-документ или полиграммной. При использовании морфологического или синтаксического анализа необходимо текстовый документ приводить к виду лингвистической разметки. Результаты синтаксического анализа также можно представить в виде деревьев зависимостей и деревьев составляющих. Семантический этап анализа на выходе формирует семантическую сеть естественного языка.

Рассмотрим подробнее способы формализации текстов с точки зрения возможности последующей автоматизированной обработки документов указанного выше вида.

При использовании вектора признаков VSM (VectorSpaceModel) текстовый документ представляется в виде вектора, каждая координата которого соответствует частоте встречаемости одного из слов всей коллекции в этом тексте. Объединение всех таких векторов в единую таблицу приводит нас к прямоугольной матрице размером $n \times p$, где p – количество слов в коллекции (размерность пространства), а n – число документов [5].

Применение полиграммной модели со степенью n и основанием M предполагает представление текстового документа в виде вектора $\{f_i\}$, $i = 1, ..., M^n$, где f_i – частота встречаемости i-й n-граммы в тексте, которая является последовательностью подряд идущих n символов вида $a_1, \ldots, a_{n-1}, a_n$, причем символы a_i принадлежат алфавиту, размер которого совпадает с M [6].

Терм-документ представляет модель, в рамках которой текст описывается лексическим вектором $\{\tau_i\}$, $i = 1, \ldots, N_w$, где τ_i – важность (информационный вес) термина w_i в документе; N_w – полное количество терминов в документной базе (словаре). Вес термина, отсутствующего в документе, принимается равным 0 [6].

Если в процессе анализа задействованы синтаксические и семантические этапы, для них необходимо текстовую информацию представлять в другом виде для сохранения результатов предыдущих этапов, а также для записи новых характеристик лингвистических единиц.

Для представления текстовых документов с синтаксическими характеристиками чаще всего используется лингвистическая разметка, предполагающая задание информации о лингвистических единицах непосредственно в тексте в форме разметки на специальном языке (например SGML или XML).

Использование специальной разметки для представления лингвистической информации в документе достаточно удобно для задач обработки текстов на естественном языке. Данный подход позволяет анализировать результаты обработки текстов пользователем или разработчиком, игнорировать не относящуюся к задаче разметку и использовать стандартные программные инструменты. Основной проблемой применения специальной разметки являются трудности при представлении сложных и пересекающихся структур, которые могут возникнуть вследствие неоднозначности анализа текста на одном из этапов обработки [7].

Грамматика зависимостей предполагает, что предложения текста можно структурировать в виде деревьев зависимостей, в которых слова связаны ориентированными дугами, обозначающими синтаксическое подчинение между главным и зависимым словами (рис. 2).

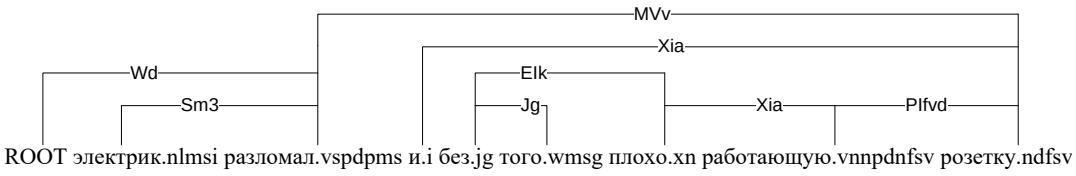

Рис. 2. Синтаксическое дерево зависимостей для предложения текста

Fig. 2. Syntax dependency tree for a text sentence

Главное отличие синтаксических деревьев зависимостей заключается в отсутствии обозначающих составляющие нетерминальных вершин, а синтаксические связи имеют пометки, обозначающие их тип. Типы связей определяют грамматические функции слов в предложении или общие семантические отношения между словами [8].

Дерево составляющих представляет собой модель формализации текста, в которой слова в предложении естественного языка группируются в составляющие на основе лингвистических наблюдений того, что цепочки слов в предложении могут функционировать как единое целое и подчиняться единым грамматическим правилам [8]. Составляющие можно перенести в середину или в конец предложения целиком, но частично перенести их без потери смысла нельзя. Пример дерева составляющих приведен на рисунке 3.

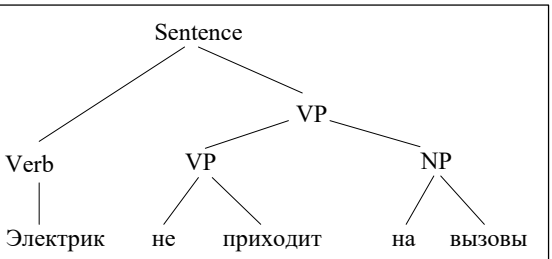

Рис. 3. Дерево составляющих предложения

Fig. 3. The tree of sentence parts

Конечная смысловая структура текста представляется основной алгебраической системой вида $M_t = \langle O_t, A_t, L_t, H_t, R_t^1, R_t^2, R_H \rangle$, называемой семантической сетью естественного языка, где O_t – множество концептов, выделенных в тексте; A_t – множество ребер, связывающих концепты из O_t; $L_t \subset L$ – множество семантических отношений, выявленных в тексте и используемых в качестве меток ребер из A_t; H_t – множество классов, связывающих концепты из O_t по классовой семантической совместимости их наборов значений семантических характеристик; R_t^1 – отношение инцидентности на $O_t \times A_t \times N$, где N – подмножество идентификаторов участников отношений модели M_2; R_t^2 – отношение инцидентности на $A_t \times L_t$; R_h – отношение классовой принадлежности на $O_t \times H_t$. Такая довольно громоздкая структура получается после нестрогого отождествления понятий из семантических образов отдельных предложений, в процессе которого образуются концепты [8].

В связи с особенностью рассматриваемых тестовых документов, поступающих на интернет-порталы органов исполнительной и законодательной власти, целесообразно использовать все перечисленные способы формализации текстов на отдельных этапах их автоматизированной обработки. Например, на этапах, предшествующих морфологическому анализу, можно использовать модель терм-документ, так как данный вид формализации очень удобен для подготовки документа к морфологическому анализу. На этапах морфологического и синтаксического анализов рационально применять также лингвистическую разметку с использованием весовых коэффициентов и экспертной информации [9], при этом необходимо учитывать нестационарность тезауруса (в том числе изменение весовых коэффициентов важности отдельных слов и их сочетаний), который зависит от выхода новых нормативных документов, выступлений должностных лиц и политических деятелей и т.д.

В процессе использования известных программных продуктов для проведения дополнительных этапов анализа придется столкнуться с проблемой разнообразия лингвистических разметок. Например, большинство синтаксических парсеров на выходе представляет каждое предложение текста в виде деревьев зависимостей, которые описывают лингвистической разметкой. Лингвистическую разметку для дальнейшей классификации и назначения весовых коэффициентов необходимо модифицировать, тем самым увеличивая размерность метрики.

Например, при использовании синтаксического парсера LinkGrammar при анализе предложения «Состояние труб водоснабжения очень плохое» получаем лингвистическую разметку:

«("LEFT-WALL" RW:6:RIGHT-WALL Wd:1:состояние.ndnsi)(Wd:0:LEFT-WALL "состояние.ndnsi" Mg:2:труб.ndfpg)(Mg:1:состояние.ndnsi "труб.ndfpg" Mg:3:водоснабжения.ndnsg)(Mg:2:труб.ndfpg "водоснабжения.ndnsg")("[очень]")("[плохое]")(RW:0:LEFT-WALL "RIGHT-WALL"))».

Дерево зависимостей данного предложения выглядит следующим образом:

```
+----------Wd--------+------Mg-----+---------Mg---------+
|          |         |             |
LEFT-WALL состояние.ndnsi труб.ndfpg водоснабжения.ndnsg [очень] [плохое]
```

Синтаксический парсер MalpParser вместо подобной XML-разметки использует построчное разбиение предложений на слова, которым приписываются характеристики в порядке, описанном в XML-файле конфигурации, который выглядит следующим образом:

```
<?xml version="1.0" encoding="UTF-8"?>
<dataformat name="conllx">
    <column name="ID" category="INPUT" type="INTEGER"/>
    <column name="FORM" category="INPUT" type="STRING"/>
    <column name="LEMMA" category="INPUT" type="STRING"/>
    <column name="CPOSTAG" category="INPUT"
type="STRING"/>
    <column name="POSTAG" category="INPUT" type="STRING"/>
    <column name="FEATS" category="INPUT" type="STRING"/>
    <column name="HEAD" category="HEAD" type="INTEGER"/>
    <column name="DEPREL"
category="DEPENDENCY_EDGE_LABEL" type="STRING"/>
    <column name="PHEAD" category="IGNORE" type="INTEGER"
default="_"/>
    <column name="PDEPREL" category="IGNORE"
type="STRING" default="_"/>
</dataformat>
```

При этом пример рассматриваемого предложения текста может выглядеть так:

```
1 They they PRON    PRP     Case=Nom|Number=Plur   2 nsubj 2:nsubj|4:nsubj
2 buy buy VERB      VBP     Number=Plur|Person=3|Tense=Pres    0 root 0:root
3 and and CONJ      CC      _                                  4 cc 4:cc
4 sell sellVERB     VBP Number=Plur|Person=3|Tense=Pres    2 conj 0:root|2:conj
5 books book NOUN NNS       Number=Plur                2 obj 2:obj|4:obj
6     .     . PUNCT .       _                          2 punct 2:punct
```

В этом случае для использования сторонних систем анализа необходимо написать программы преобразования форматов представления текстовой информации, а также модифицировать предложенные форматы для внесения дополнительных характеристик, необходимых для модифицируемых методов классификации текстовых документов. Например, для жалобы «Налоговая инспекция продолжает уже два месяца кошмарить нашу фирму» синтаксический парсер LinkGrammar представит его в лингвистической разметке вида:

(("LEFT-WALL" RW:10:RIGHT-WALL Wd:3:продолжает.vnpdn3s)("налоговая.afsi" Afi:2:инспекция.ndfsi)(Afi:1:налоговая.afsi "инспекция.ndfsi" Sf3:3:продолжает.vnpdn3s)(Wd:0:LEFT-WALL Sf3:2:инспекция.ndfsi "продолжает.vnpdn3s" I:7:кошмарить.vsndi E:4:уже.as)(E:3:продолжает.vnpdn3s "уже.as")("два" IDBBT:6:месяца)(IDBBT:5:два "месяца" EI:7:кошмарить.vsndi)(I:3:продолжает.vnpdn3s EI:6:месяца "кошмарить.vsndi" MVv:9:фирму.ndfsv) ("нашу.wfsv" Afv:9:фирму.ndfsv)(MVv:7:кошмарить.vsndi Afv:8:нашу.wfsv "фирму.ndfsv") (RW:0:LEFT-WALL "RIGHT-WALL")).

С точки зрения синтаксиса важными словами являются «налоговая», «инспекция» и «бизнес» – они будут помечены как синтаксически важные, и им присвоится специальный коэффициент. Также в системе из-за динамичности тезауруса, который связан с нормативными документами и с выступлениями должностных лиц, будет учитываться слово «кошмарить», которое относится к проблемам малого бизнеса, так как именно в данном контексте его использовал один из руководителей страны. Учитывая все названные замечания, модифицированная лингвистическая разметка будет выглядеть следующим образом:

(("LEFT-WALL" RW:10:RIGHT-WALL Wd:3:продолжает.vnpdn3s.com.n_imp)("налоговая.afsi.uni.imp" Afi:2:инспекция.ndfsi)(Afi:1:налоговая.afsi "инспекция.ndfsi.rare.imp" Sf3:3:продолжает.vnpdn3s)(Wd:0:LEFT-WALL Sf3:2:инспекция.ndfsi "продолжает.vnpdn3s.com.n_imp" I:7:кошмарить.vsndi E:4:уже.as)(E:3:продолжает.vnpdn3s "уже.as.com.m_imp")("два.com.n_imp" IDBBT:6:месяца)(IDBBT:5:два "месяца.com.n_imp" EI:7:кошмарить.vsndi)(I:3:продолжает.vnpdn3s EI:6:месяца "кошмарить.vsndi.uni.n_imp.normative" MVv:9:фирму.ndfsv)("нашу.wfsv.com.n_imp" Afv:9:фирму.ndfsv)(MVv:7:кошмарить.vsndi Afv:8:нашу.wfsv "фирму.ndfsv.uni.imp")(RW:0:LEFT-WALL "RIGHT-WALL")).

В разметке указаны общие слова – com, редкие – rare и уникальные – uni, а также стоят пометки важности – imp или n_imp, и, если слова относятся к нормативным высказываниям, normative. Усовершенствовав разметку, можно применять метод рубрицирования, изложенный в [9].

Сказанное позволяет формализовать процедуру обработки текстовых документов при рубрицировании текстов в автоматизированных системах органов государственного и муниципального управления (рис. 4).

Для проверки целесообразности использования синтаксических коэффициентов важности слов в предложении был проведен эксперимент, описанный в работе [10], но с дополнительным использованием синтаксических парсеров. В эксперименте были рубрицированы 100 документов по 9 областям, при этом среднее количество слов в документе составляло 28. Анализ включал этапы сегментации, морфологического анализа, сбора статистических данных, рубрицирование (классификация) текстового документа. При учете синтаксических коэффициентов важности дополнительно добавлялись разметки весовых коэффициентов синтаксической важности слов и сбора статистических данных. Под сегментацией понималось разбиение текстового документа на абзацы, предложения и слова; под морфологическим анализом – нахождение лингвистических характеристик всем словам, таких как начальная форма слова, часть слова, род, число, падеж, форма и т.д.; под сбором статистических данных – подсчет частотных характеристик всех слов; под синтаксическим анализом – построение дерева зависимостей для всех предложений текстового документа; под разметкой весовых коэффициентов синтаксической важности слов – назначение коэффициентов исходя из синтаксической значимости слов в предложениях; под рубрицированием (классификацией) – подсчет степени принадлежности текстового документа ко всем предметным областям по математической модели метода, описанного выше, и нахождение максимума. Программирование этапов анализа и преобразования форматов представления текстовых документов осуществлялось на языке Micro-

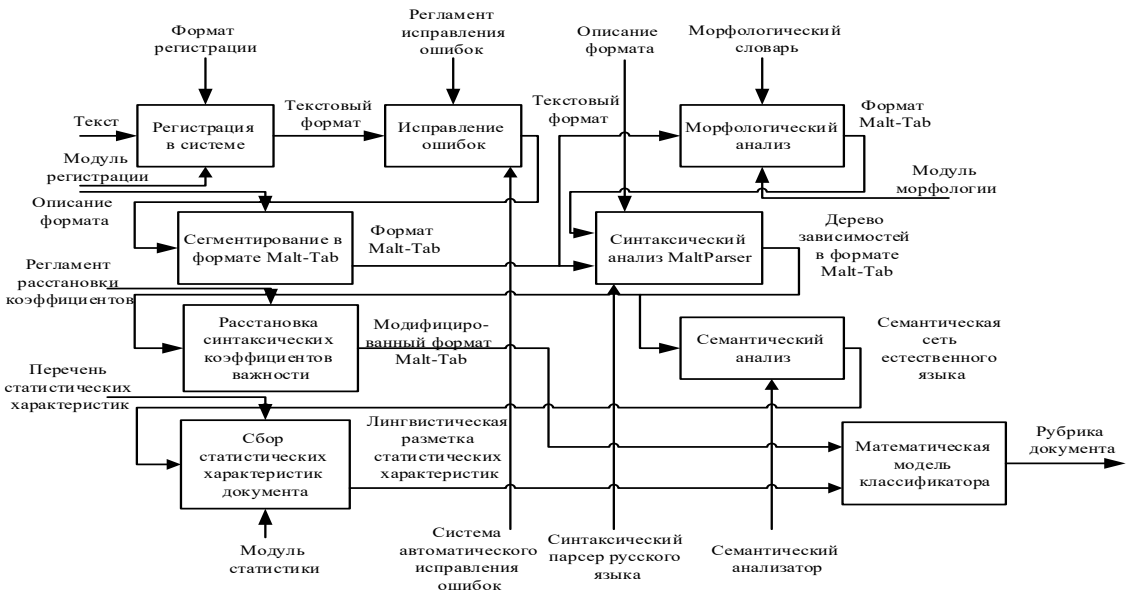

Рис. 4. Диаграмма процедуры обработки текстовых документов при рубрицировании текстов в автоматизированных системах органов государственного и муниципального управления

Fig. 4. The diagram of a text document processing procedure for text rubrication in automated systems of state and municipal government

soft Visual C# под управлением ОС Microsoft Windows 8.1. БД хранятся в сетевом экземпляре Microsoft SQL Server 2008.

При проведении эксперимента параметры весового метода были следующие: уникальным словам приписывается вес 50, неуникальным – 10, общим – 1, порог отбора общих слов – 80 %, а синтаксические коэффициенты важности: подлежащее – 10, сказуемое – 10, остальные – 1.

Наблюдаемое увеличение точности автоматизированного рубрицирования до 84 % подтвердило целесообразность использования синтаксических парсеров для анализа.

Предполагается, что более тщательный подбор коэффициентов синтаксической важности слов и проверка синтаксических связок слов могут повысить точность приведенной на рисунке 1 процедуры. Очевидно, что при нестационарности тезауруса целесообразно для определения степени важности отдельных слов и их сочетаний применять экспертные процедуры и интеллектуальные системы, в том числе с использованием аппарата теории нечетких множеств [11–13], а также методы когнитивного моделирования [14–16].

Литература

1. Аналитическая справка о работе Аппарата администрации Смоленской области с обращениями граждан. URL: https://www.admin-smolensk.ru/obrascheniya_grazhdan/obzori_obrascheniy/news_16096.html (дата обращения: 20.03.2017).

2. Обзор обращений граждан Администрации города Санкт-Петербурга. URL: http://gov.spb.ru/gov/obrasheniya-grazhdan/otchet-obrasheniya/?page=1 html (дата обращения: 20.03.2017).

3. Дли М.И., Какатунова Т.В. О перспективах создания виртуальных технопарковых структур // Инновации. 2008. № 2. С. 118–120.

4. Дли М.И., Какатунова Т.В. Общая процедура взаимодействия элементов инновационной среды региона // Журнал правовых и экономических исследований. 2009. № 3. С. 60–63.

5. Учителев Н.В. Классификация текстовой информации с помощью SVM // Информационные технологии и системы. 2013. № 1. С. 335–340.

6. Андреев А.М., Березкин Д.В., Морозов В.В., Симаков К.В. Автоматическая классификация текстовых документов с использованием нейросетевых алгоритмов и семантического анализа // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: тр. V Всерос. науч. конф. (RCDL'2003). СПб, 2003. С. 140–149.

7. Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика. М.: Изд-во МИЭМ, 2011. 272 с.

8. Шелманов А.О. Исследование методов автоматического анализа текстов и разработка интегрированной системы семантико-синтаксического анализа: дисс. … канд. технич. наук. М., 2015. 210 с.

9. Козлов П.Ю. Методы автоматизированного анализа коротких неструктурированных текстовых документов // Программные продукты и системы. 2017. № 1. С. 100–105.

10. Козлов П.Ю. Сравнение частотного и весового алгоритмов автоматического анализа документов // Науч. обозрение. 2015. № 14. С. 245–250.

11. Круглов В.В., Дли М.И., Голунов Р.Ю. Нечеткая логика и искусственные нейронные сети. М.: Физматлит, 2001.

12. Круглов В.В., Дли М.И. Интеллектуальные информационные системы // Компьютерная поддержка систем нечеткой логики и нечеткого вывода. М.: Физматлит, 2002. 256 с.

13. Федулов А.С. Устойчивая операция аккумулирования нечетких чисел // Нейрокомпьютеры: разработка, применение. 2007. № 1. С. 27–39.

14. Дли М.И., Какатунова Т.В. Нечеткие когнитивные модели региональных инновационных систем // Интеграл. 2011. № 2. С. 16–18.

15. Borisov V.V., Fedulov A.S. Generalized rule-based fuzzy cognitive maps: structure and dynamics model. LNCS, 2004, vol. 3316, pp. 918–922.

16. Дли М.И., Какатунова Т.В. Функциональные когнитивные карты для моделирования региональных инновационных процессов // Инновационная деятельность. 2011. № 3. С. 75–83.

Software & Systems Received 21.03.17
DOI: 10.15827/0236-235X.120.678-683 2017, vol. 30, no. 4, pp. 678–683

METHODS OF REPRESENTING TEXT INFORMATION IN AUTOMATED RUBRICATION OF SHORT TEXT DOCUMENTS

P.Yu. Kozlov [1], *Postgraduate Student, originaldod@gmail.com*

[1] *Smolensk Branch of the Moscow Power Engineering Institute, Energeticheskiy proezd 1, Smolensk, 214013, Russian Federation*

Abstract. The paper shows that citizens' electronic messages (complaints, appeals, proposals, etc.) in terms of the possibility of their automated processing have a number of specific features. They are: usually a small document capacity, which makes it difficult to analyze it statistically, a lack of structuring, which complicates extracting information, a big number of grammatical and syntactic errors that lead to implementing several additional processing steps, thesaurus non-stationarity (composition and importance of words), which depends on the issuance of new normative documents, officials' and politicians' speeches, etc. All this leads to the necessity of using procedures for headings dynamic classification.

The paper describes the stages of automated analysis and methods for formalizing text documents. It also proposes a developed rubrication method that uses the results of the morphological and syntactic stages with modified linguistic markup of text documents. The syntactic parser is MaltParser or LinkGrammar software that build dependency trees for all sentences in a document. The paper shows standard linguistic markings of MaltParser and LinkGrammar applied to short text documents, as well as a modification of the LinkGrammar markup to use for rubrication. Using known software for additional stages of analysis shows the problem of the diversity of linguistic markings. For example, most of the syntactic parsers at the output represent each sentence as dependency trees, which are described by linguistic markup. For further classification and assignment of weighting factors, linguistic markup should be modified, so it will increase the dimension of the metric.

The developed method of rubrication takes into account the expert evaluation of the importance of words for each rubric, as well as the syntactic role of words in sentences. The paper shows a diagram of the process of automated rubrication of complaints and proposals in the developed analysis system. It also describes an experiment that confirms the expediency of using syntactic parsers in such systems, which leads to increasing accuracy of rubrication.

There are recommendations to improve the accuracy of the developed method and use the theory of fuzzy sets and methods of cognitive modeling in order to solve the problem of thesaurus nonstationarity in the systems that depend on the issue of normative documents and officials' speeches.

Keywords: automated text analysis, dynamic thesaurus, methods for formalizing text documents.

References

1. *Analiticheskaya spravka o rabote Apparata administratsii Smolenskoy oblasti s obrascheniyami grazhdan* [An Analytical Report on the Work of the Smolensk Region Administration with Citizens' Applications]. Available at: https://www.admin smolensk.ru/obrascheniya_grazhdan/obzori_obrascheniy/news_16096.html (accessed March 20, 2017).

2. *Obzor obrascheny grazhdan Administratsii goroda Sankt-Peterburga* [An Overview of Citizens' Applications by St. Petersburg Administration]. Available at: http://gov.spb.ru/gov/obrasheniya-grazhdan/otchet-obrasheniya/?page=1 (accessed March 20, 2017).

3. Dli M.I., Kakatunova T.V. On the prospects of creating virtual technopark structures. *Innovatsii* [Innovations]. 2008, no. 2, pp. 118–120 (in Russ.).

4. Dli M.I., Kakatunova T.V. General procedure of interaction between elements of region innovation environment. *Zhurnal pravovykh i ekonomicheskikh issledovany* [Journal of Legal and Economic Studies]. 2009, no. 3, pp. 60–63 (in Russ.).

5. Uchitelev N.V. Classification of text information using SVM. *Informatsionnye tekhnologii i sistemy* [Information Technologies and Systems]. 2013, no. 1, pp. 335–340 (in Russ.).

6. Andreev A.M., Berezkin D.V., Morozov V.V., Simakov K.V. Automatic text classification using neuronets algorithms and semantic analysis. *Tr. Konf. RCDL 2003* [Proc. RCDL 2003 Conf.]. Available at: http://rcdl.ru/doc/2003/B1.pdf (accessed November 1, 2017).

7. Bolshakova E.I., Klyshinsky E.S., Lande D.V., Noskov A.A., Peskova O.V., Yagunova E.V. *Avtomaticheskaya obrabotka tekstov na estestvennom yazyke i kompyuternaya lingvistika* [Automatic Text Processing in Natural Language and Computer Linguistics]. Moscow, MIEM Publ., 2011, 272 p.

8. Shelmanov A.O. *Issledovanie metodov avtomaticheskogo analiza tekstov i razrabotka integrirovannoy sistemy semantiko-sintaksicheskogo analiza* [Research on Methods of Automatic Text Analysis and the Development of an Integrated System of Semantic-Syntactic Analysis]. Ph.D. diss. FITs IU RAN Publ., Moscow, 2015, 210 p.

9. Kozlov P.Yu. Automated analysis method of short unstructured text documents. *Programmnye produkty i sistemy* [Software & Systems]. 2017, no. 1, pp. 100–105 (in Russ.).

10. Kozlov P.Yu. Comparison of frequency and weight algorithms for automatic document analysis. *Nauchnoe obozrenie* [Science Review]. 2015, no. 14, pp. 245–250 (in Russ.).

11. Kruglov V.V., Dli M.I., Golunov R.Yu. *Nechetkaya logika i iskusstvennye neyronnye seti* [Fuzzy Logic and Artificial Neural Networks]. Moscow, Nauka, Fizmatlit Publ., 2001.

12. Kruglov V.V., Dli M.I. *Intellektualnye informatsionnye sistemy: kompyuternaya podderzhka sistem nechetkoy logiki i nechetkogo vyvoda* [Intellectual Information Systems: Computer Support of Fuzzy Logic and Fuzzy Inference Systems]. Moscow, Fizmatlit Publ., 2002, 256 p.

13. Fedulov A.S. Stable operation of fuzzy numbers accumulation. *Neyrokompyutery: razrabotka, primenenie* [Journal Neurocomputers]. 2007, no. 1, pp. 27–39 (in Russ.).

14. Dli M.I., Kakatunova T.V. Fuzzy cognitive models of regional innovation systems. *Integral*. 2011, no. 2, pp. 16–18 (in Russ.).

15. Borisov V.V., Fedulov A.S. Generalized rule-based fuzzy cognitive maps: structure and dynamics model. *Lecture Notes in Computer Science*. 2004, vol. 3316, pp. 918–922 (in Russ.).

16. Dli M.I., Kakatunova T.V. Functional cognitive maps for modeling regional innovation processes. *Innovatsionnaya deyatelnost* [Innovative Activity]. 2011, no. 3, pp. 75–83 (in Russ.).