КОМПЬЮТЕРНАЯ ЛИНГВИСТИКА И ОБРАБОТКА ЕСТЕСТВЕННОГО ЯЗЫКА

УДК 004.912

ПРИМЕНЕНИЕ КОРПУСА ТЕКСТОВ ДЛЯ АВТОМАТИЧЕСКОЙ КЛАССИФИКАЦИИ В КОМПЛЕКСЕ ИНСТРУМЕНТОВ АВТОМАТИЗИРОВАННОГО АНАЛИЗА ТЕКСТОВ

С. А. Полицын, Е. В. Полицына

Московский авиационный институт (национальный исследовательский университет)

Поступила в редакцию 29.01.2018 г.

Аннотация. Одной из актуальных задач компьютерной лингвистики, решаемой в рамках комплекса инструментов автоматизированного анализа текстов, является автоматическая классификация текстов. Для обучения классификатора на большом наборе предметных областей актуальной является задача полной автоматизации этого процесса, что требует наличия размеченного корпуса текстов. В статье описывается создание корпуса текстов с расширяемому разметкой и приложения для работы с ним, которое позволяет создавать субкорпуса по настраиваемому набору признаков. Это дает возможность использоваться корпус, как для обучения при решении других задач анализа текста, так и для автоматизации проверки получаемых результатов при исследовании различных методов компьютерной лингвистики.

Ключевые слова: корпус текстов, классификация текстов, инструменты автоматизированного анализа текстов, обучение классификатора.

Annotation. One of the urgent tasks of computer linguistics is automatic classification of texts. This task is solved in the complex of tools for automated text analysis developed by authors too. To train the classifier on a large set of subject areas, it is necessary to automate this process, which requires the presence of a marked textual corpus. The article describes the creation of the corpus of texts with extensible markup and an application for working with it, which allows creating subcorpora according to a custom set of characteristics. This allows using the corpus both for machine-learning methods training during solving tasks of text analysis, and for automating the verification of results of various methods of computer linguistics.

Keywords: automated text analysis tools, corpus of texts, classifier training.

ВВЕДЕНИЕ

С бурным ростом количества обрабатываемой информации последние десятилетия потребность в развитии методов и инструментов компьютерной лингвистики только увеличивается. Одной из задач компьютерной лингвистики является автоматическая классификация текстов, т. е. отнесение текста к той или иной области или ее подмножеству на основе некоторого алгоритма с некоторой вероятностью.

Часть алгоритмов используют для этого только данные, полученные непосредственно из этого текста, такие алгоритмы имеют невысокую точность и часто не соответствуют решению задачи классификации человеком, часть алгоритмов использует дополнительную информацию (обучающие выборки текстов, словари предметных областей, списки слов-признаков и т. д.), что требует подготовки дополнительных данных. Начиная с ранних этапов развития, отдельный субъект познания развивает свое представление об окружающем мире, приобретая новые знания, и с каждым разом всё лучше справляет-

ся с задачей классификации текста, а будучи специалистом в какой-либо более узкой области может максимально точно классифицировать текст.

Поэтому любая компьютерная система классификации должна быть самообучающейся, не зависимо от используемого алгоритма классификации, с каждым разом решая задачу всё точнее, используя весь накопленный за время работы опыт.

В комплексе инструментов автоматизированного анализа текстов [1] реализованы инструменты анализа и исследования текстов на этапах морфологического, синтаксического анализа, с применением статистических методов, кроме того присутствует средство исследования полученных результатов на следующем – аналитическом – уровне. На основе инструментов комплекса созданы сервисы решения задач выделения ключевых слов, статистического анализа, классификации, представленные на портале «Автоматизированный анализ текста» [2].

Сервис классификации текстов имеет два режима работы: режим анализа и режим обучения, – в основе его работы с ключевыми словами, полученными в результате расчета частоты употребления слов в тексте с применением морфологического анализа и средств аналитической обработки [3]. Внедрение других методов классификации позволит улучшить точность результатов, но анализ результатов классификации показал существенное увеличение качества получаемых результатов после обучения классификатора [3] в рамках одной предметной области. Для обучения классификатора был подготовлен набор текстов, которые были вручную классифицированы и загружены в систему средствами сервиса классификации. Для обучения классификатора на большом наборе предметных областей актуальной является задача полной автоматизации этого процесса, что требует наличия размеченного корпуса текстов.

ОБЗОР КОРПУСОВ ТЕКСТОВ

Существует множество классификаций корпусов текстов: по способу их построения (статические и динамические, одноязычные и многоязычные, размеченные и неразмеченные и т. д.), распространения (свободно или частично доступные, закрытые), назначению и т. д. [4, 5].

В зависимости от решаемой задачи возникает необходимость в наличии различной информации о текстах в корпусе [4, 6]: морфологической, синтаксической, семантической. Например, для решения задачи определения тональности текста необходимо указание частей речи слов, разметка предложений, семантическая разметка и прочее. Для обучения системы классификации текстов требуется разметка корпуса текстов в соответствии с выбранными признаками классифицирования (стиль текста, жанр, автор, тематика, дата написания и т. д.), кроме того корпус текстов должен быть достаточно большим, чтобы обеспечивать репрезентативность выборки и приемлемое качество обучения классификатора, в том числе при увеличении количества единиц разбиения (бинарная, многокритериальная или фасетная классификации).

Среди существующих размеченных корпусов русскоязычных текстов были проанализированы корпуса текстов, доступные в соответствии с некоммерческой лицензией: «Национальный корпус Русского языка» [7], Хельсинкский аннотированный корпус русских текстов ХАНКО [8], «Генеральный Интернет-Корпус Русского Языка» [9], «Открытый корпус» [10]. Эти корпуса содержат большой объем данных, многие содержат максимально полную морфологическую разметку, которая полезна для решения других задач компьютерной лингвистики, но в большинстве отсутствует разметка, необходимая для классификации текстов и тем более возможность ее настройки. Существуют онлайн варианты корпусов, но они не очень пригодны для автоматического обучения классификатора в связи с ограничением списка возвращаемых результатов и закрытостью исходных корпусов текстов.

Одним из популярных направлений в компьютерной лингвистике является определение тональности текста или отзыва о чем-либо (например, [11]), большая часть

известных алгоритмов решения этой задачи опирается на методы машинного обучения, и, как следствие, при применении данных методов требуется подготовка корпусов текстов [6]. Часто имея достаточный объем, они предназначены только для обучения классификатора для определения тональности текстов на основе ручного анализа подборки текстов сервиса микроблогов Twitter.

Таким образом, необходима разработка корпуса текстов с разметкой, пригодной для обучения классификатора текстов вне зависимости от способа его реализации (использование статистических методов компьютерной лингвистики, машинного обучения, нейронных сетей). Такой корпус по возможности должен быть универсальным: пригодным для обучения классификаторов по различным пополняемым наборам признаков и с возможностью создания на его основе субкорпусов для решения той или иной задачи.

СОЗДАНИЕ КОРПУСА

Для создания корпуса текстов в первую очередь требуется решение двух задач: получение исходных текстовых данных в достаточном количестве (порядка нескольких терабайт всего или около тысячи текстов среднего размера на единицу классификации) и разработка языка разметки для них для автоматического обучения классификатора.

Среди художественных текстов в сети Интернет доступна подборка произведений авторов XVIII–XX веков; тексты современных авторов, а также большие подборки публицистических и научных текстов по большей части отсутствуют в свободном доступе, доступны ограниченные наборы газетных статей.

Для получения достаточного объема нехудожественных текстов по разным областям за основу была взята коллекция рефератов (bankreferatov.ru), дополненная доступными научными статьями (elibrary.ru), общий объем собранной коллекции в настоящее время около 23 Гб.

Решение задачи обучения классификатора по различным областям также требует разработки системы разметки корпуса, позволяющей иметь неограниченное число признаков классификации. Для различных целей требуются выборки по тематикам, авторам, поднаправлениям внутри направления и т. д. При недостаточной выборке по какому-либо признаку возникает потребность в расширении корпуса. Этого можно достигнуть путем применения какого-либо вида разметки для исходного текста, в которой будут в частично произвольном виде указываться характеристики данного текста. Поэтому, оставляя возможность ручного создания выборки, лучшим вариантом разметки является стандартизованная разметка на основе языка XML. Основной недостаток XML-разметки – это ее избыточность, однако при хранении больших корпусов размеченных текстов использование сжатия позволит устранить этот недостаток. В целом же удобная работа с XML-разметкой поддерживается всеми современными языками программирования и оставляет возможность редактирования человеком.

Только теги **<doc>** и **<text>** являются обязательным, остальные могут быть добавлены по мере наполнения корпуса, например, теги **<category>**, **<author>**, **<title>**, **<keywords>**:

<doc> – корневой тег

<category> – категория, дерево через / (Пример: Медицина/Стоматология)

<author> – автор

<title> – название,

<keywords> – ключевые слова автора

<text><![CDATA[ТЕКСТ]]></text> – текст

На рис. 1 представлен пример XML-документа с фрагментом текста и указанной предметной областью.

Для удобства работы с корпусом было создано приложение, позволяющее произвести выборку текстов по имеющимся тегам. Расширение набора тегов не ограничено и будет воспринято программой как дополнительные заданные атрибуты текста. Пользователь с помощью графического интерфейса приложения может отобрать тексты по тем из доступных тегов, которые необходимы ему для решаемой задачи (рис. 2).

В дальнейшем, планируется создание веб-сервиса на его основе для получения

```
<?xml version="1.0" ?>
<doc>
  <category>Культура и искусство</category>
  <text>
<![CDATA[Своеобразие индийской религии: брахманизм, индуизм, тантризм.
Брахманизм может рассматриваться как ранняя стадия индуизма, главное
содержание которой состоит в переходе от ведийской мифологии и политеизма
(ведическая религия) к идее единого бога-творца и затем — абстрактного абсолюта,
основы и причины мира. Этот переход начинается уже в «Ригведе».
В некоторых гимнах, особенно поздних, выражается неудовлетворенность архаической ,
...
имеет в тантризме глубокий смысл. ]]>
</text>
</doc>
```

Рис. 1. Пример XML-документа корпуса текстов

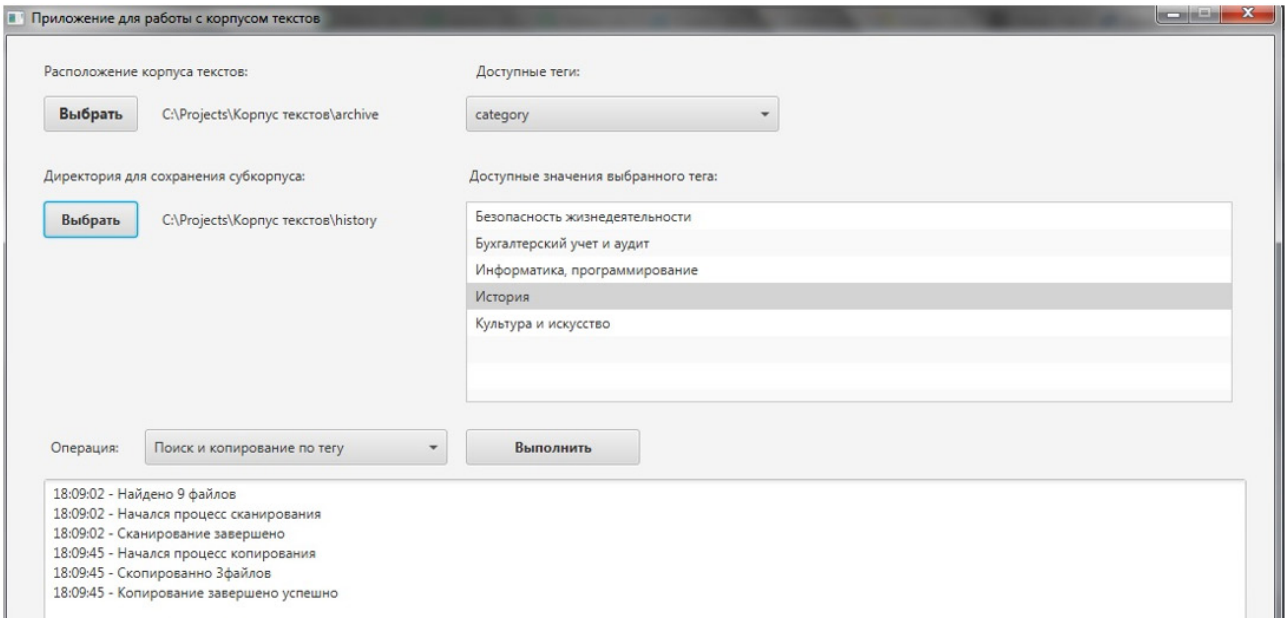

Рис. 2. Интерфейс приложения для работы с корпусом

необходимых исследователям субкорпусов. Приложение позволяет на основе корпуса создавать субкорпуса, ориентированные на решение отдельной конкретной задачи [12].

ПРИМЕНЕНИЕ КОРПУСА В КОМПЛЕКСЕ ИНСТРУМЕНТОВ АВТОМАТИЗИРОВАННОГО АНАЛИЗА ТЕКСТОВ

В настоящее время создаваемый корпус наполняется текстами с указанием предметных областей, которые выстроены в иерархию, авторов, стиля текста, его объема.

Разработанная структура корпуса и приложение для работы с ним позволяют применять его не только при решении задачи классификации текстов, но и анализа работы различных методов (например, выделения ключевых слов и словосочетаний) в зависимости стиля текста, его объема и т. д., больше автоматизировать проведение исследований текстов произведений разных авторов и др.

Возможность добавления произвольных дополнительных тегов позволяет готовить не только обучающие субкорпуса, но и контрольные – для автоматизированной проверки качества работы разных методов анализа текста.

Для хранения корпуса текстов используется сервис Яндекс.Диск. Автоматическая генерация имени файла с текстом позволяет

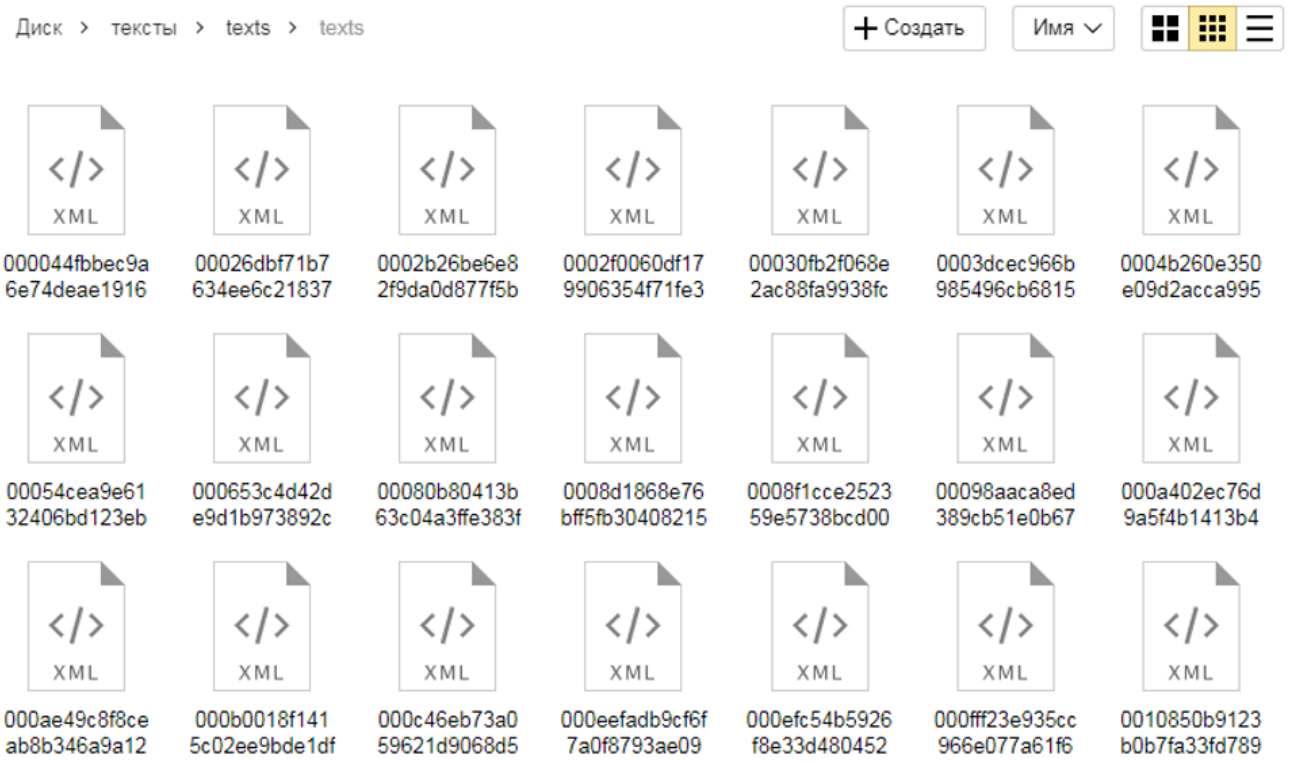

Рис. 3. Хранение корпуса текстов

обсепечить их уникальность, тогда как вся необходимая информация о тексте содержится в самом файле (рис. 3).

Разработанный корпус применяется в комплексе инструментов автоматизированного анализа текстов для исследования методов выделения ключевых слов и словосочетаний, классификации текстов, получения кратких изложений текстов (аннотаций и рефератов) [13–15].

ЗАКЛЮЧЕНИЕ

Создание общедоступного корпуса текстов и приложения для работы с ним позволяет автоматизировать процесс обучения разрабатываемого классификатора в комплексе инструментов автоматизированного анализа текстов и может быть полезным при формировании субкорпусов, ориентированных на решение задачи обучения классификатора методом машинного обучения.

Расширяемая система тегов и разработанное приложение для создания субкорпусов по настраиваемому набору признаков позволяет использоваться корпус, как для обучения при решении других задач анализа текста, так и для автоматизации проверки получаемых результатов при исследовании различных методов компьютерной лингвистики.

Разработка веб-сервиса позволит реализовать возможность свободного распространения не только корпуса целиком, но и получение нужного пользователю субкорпуса для проведения исследований или решения практических задач анализа текста.

СПИСОК ЛИТЕРАТУРЫ

1. *Балакирев Н. Е., Полицына Е. В.* Подход к созданию комплекса инструментов автоматизированного анализа текстов на русском языке // Вестник Воронеж. гос. ун-та. Сер. Системный анализ и информационные технологии. – 2016. – № 2. – С. 98–105.

2. Портал «Автоматизированный анализ текста» [Электронный ресурс]. Режим доступа: http://textanalysis.ru/ (Дата обращения: 23.02.2018).

3. *Полицына, Е. В.* Создание настраиваемого сервиса классификации в составе открытой системы автоматизированного анализа текста // Материалы XIII Международной научно-методической конференции «Информа-

тика: проблемы, методология, технологии». Т. 1. Воронеж, 2013. – С. 73–77.

4. *Козлова, Н. В.* Лингвистические корпуса: определение основных понятий и типология / Н. В. Козлова // Вестник НГУ, Лингвистика и межкультурная коммуникация. – 2013. – Т. 11, выпуск 1. – С. 79–88.

5. *Гаинева, И. Ф.* Об использовании корпусов в лингвистических исследованиях / И. Ф. Ганиева // Вестник Башкирского университета. – 2007. – Т. 4, № 12. – С. 104–106.

6. *Рубцова Ю. В.* Построение корпуса текстов для настройки тонового классификатора / Ю. В. Рубцова // Программные продукты и системы. –2015. – № 1. – С. 72–78.

7. Национальный корпус русского языка [Электронный ресурс] – Режим доступа: http://ruscorpora.ru (Дата обращения 09.10.2017).

8. Ханко-хельсинкский аннотированный корпус [Электронный ресурс] – Режим доступа: http://www.ling.helsinki.fi/projects/hanco/ (Дата обращения 09.10.2017).

9. General Internet-Corpus of Russian [Электронный ресурс] – Режим доступа:http://www.webcorpora.ru (Дата обращения 09.10.2017).

10. «Открытый корпус» (OpenCorpora) [Электронный ресурс] – Режим доступа: http://opencorpora.org/ (Дата обращения 09.10.2017).

11. Корпус коротких текстов на русском языке на основе постов twitter [Электронный ресурс] – Режим доступа: http://study.mokoron.com/ (Дата обращения 09.10.2017).

12. Применение современных методов корпусной лингвистики при анализе текста // Молодежный научный форум: Гуманитарные науки: электр. сб. ст. по материалам XXV студ. междунар. заочной науч.-практ. конф. – М. : «МЦНО». – 2015. – № 6(24) [Электронный ресурс]. – Режим доступа: http://nauchforum.ru/archive/MNF_humanities/6(24).pdf

13. *Иващенко, М. В.* Анализ методов автоматизированного выделения ключевых слов из текстов на естественном языке / М. В. Иващенко // Материалы XVIII Международной научно-методической конференции «Информатика: проблемы, методология, технологии». Т. 6, Воронеж. – 2018. – С. 19–24.

14. *Пряженцева, А. А.* Анализ методов автоматизированного выделения ключевых слов из текстов на естественном языке / А. А. Пряженцева // Материалы XVIII Международной научно-методической конференции «Информатика: проблемы, методология, технологии». Т. 6, Воронеж. – 2018. – С. 56–60.

15. *Белов, С. М.* Создание программной системы классификации текстов / С. М. Белов // Материалы XVIII Международной научно-методической конференции «Информатика: проблемы, методология, технологии». Т. 6, Воронеж. – 2018. – С. 8–12.

Полицын Сергей Александрович – ст. преподаватель, институт № 3, кафедра 319, Московский авиационный институт (Национальный исследовательский университет).
Тел.: 8-499-141-94-82
E-mail: pul_forever@mail.ru

Полицына Екатерина Валерьевна – канд. техн. наук, доцент, институт № 3, кафедра 319, Московский авиационный институт (Национальный исследовательский университет).
Тел.: 8-499-141-94-82
E-mail: kathrin.beaver@mail.ru

Politsyn Sergey A. – senior lecturer, department of «Design of Computing Systems», Moscow Aviation Institute (National Research University).
Tel.: 8-499-141-94-82
E-mail: pul_forever@mail.ru

Politsyna Ekaterina V. – candidate of technical sciences, associate professor, department of «Design of Computing Systems», Moscow Aviation Institute (National Research University).
Tel.: 8-499-141-94-82
E-mail: kathrin.beaver@mail.ru