Программная реализация автоматизированного аннотирования документов

О.В. Свиридова, Н.Н. Короткова, А.А. Рыбанов

Волжский политехнический институт (филиал) Волгоградского государственного технического университета, г. Волжский Волгоградской обл.

Аннотация: В статье приведен обзор методов автоматического реферирования и аннотирования документов. Представлено описание классификации методов аннотирования и реферирования по различным критериям. Подробно описан алгоритм автоматического аннотирования, ключевым шагом которого является преобразование всех слов в тексте в леммы (лемматизация). В разработанной программе установлена верхняя граница объёма выводимой аннотации для того, чтобы избежать получения реферата вместо аннотации.
Ключевые слова: аннотирование, реферирование, лемматизация, алгоритм, программная реализация, база данных, анализ текста

Введение

Ежедневно возрастающий объем информации объясняет особое внимание, уделяемое проблеме автоматического аннотирования и реферирования текста.

Под аннотированием понимают краткое описание основных тем текста. Реферирование – это извлечение из текста основного содержания и необходимой информации для последующего изложения.

С появлением ЭВМ стали применять компьютеры для машинного перевода и, одновременно, автоматического реферирования текста. Первоначально для этого использовались статистические подходы. Затем начали исследовать внутренние структуры текста и его информационной основы.

Постановка проблемы

Тарасов С.Д. предлагает следующую классификацию [1] методов автоматического реферирования:

1 По типу получаемого реферата можно выделить следующие подходы.

1.1 Экстракция. Извлекаются предложения, которые затем собираются в реферат, обычно не связный. Предложения отбираются на основе позиции в тексте и ключевым словам. В методах данного подхода обязательно используется оценочная функция. Оценочная функция каждого модуля текста зависит от его расположения в исходном тексте, частоте появления, частоты использования в ключевых предложениях, и также индексов статистического значения.

1.2 Абстракция – на основе лингвистического сжатия или с опорой на знания. Происходит в три этапа: сначала происходит анализ исходного текста с генерацией внутреннего представления, затем – семантическое сжатие полученного внутреннего представления и, наконец, синтез реферата.

2. Не менее важной является классификация методов автоматического реферирования по уровню анализа исходного текста.

2.1 Поверхностный уровень. В методах этого уровня текст рассматривается как линейная структура из предложений и слов (или словосочетаний). К классическим методам этого уровня относятся статистические (используется частота встречаемости слова в тексте), позиционные методы (информативность предложения от его позиции в тексте) и индикаторные методы (для идентификации фрагментов используются слова-маркеры). К классическим методам относятся модели на основе Марковских цепей и алгоритм LRU-K, который является усовершенствованием алгоритма «Последний недавно использованный». Вторая группа методов этого уровня – методы на основе машинного обучения.

2.2 Уровень сущностей текста. Подходы этого уровня используют модель структурной связности текста.

2.3 Уровень дискурсной структуры текста. Модели основаны на анализе содержательной модели связности текста.

3. По критерию использования опоры на знания.

3.1. Методы без опоры на знания не предполагают создания специальных баз знаний какой-либо конкретной предметной области.

3.2. Методы с опорой на знания – предполагают создание баз знаний с наборами правил и эвристик конкретной предметной области.

4. По технологии построения реферата можно выделить:

4.1 Подходы «сверху-вниз» – получают внутреннее представление исходного текста, затем его преобразовывают и формируют реферат.

4.2 Подходы «снизу-вверх» – выделяются релевантные фрагменты из исходного текста и из них уже формируется реферат.

5. По ориентации на предметную область методы делятся на:

5.1. Подходы без ориентации на предметную область (Domain-independent approaches).

5.2. Подходы с ориентацией на конкретную предметную область.

Семантические методы подготовки рефератов могут использовать синтаксический разбор предложений, при этом используют деревья разбора текста, перегруппировывая их и сокращая ветви. Также к синтаксическим относятся методы, опирающиеся на понимание естественного языка, когда используют системы искусственного интеллекта. В этом случае семантические структуры хранятся в базе знаний, которые затем преобразовываются для получения реферата.

Четкой классификации методов не существует, поэтому часто метод можно отнести к разным группам, также ученые путают разные подходы [2].

Методы автоматического реферирования могут быть широко применены для обработки больших объемов информации.

Сравнительный анализ программных продуктов для аннотирования документов (ОРФО 5.0, Либретто, МедиаЛингва Аннотатор SDK 1.0, Oracle Context, Intelligent Text Miner) по методу Саати [3, 4] позволил оценить их

функциональные качества такие, как объем вводимых данных, количество типов расширения файла, количество возможных диапазонов для конвертирования, качество логического смысла получаемой аннотации, а также удобство интерфейса. Данное исследование показало, что у всех рассмотренных программных средств имеются показатели, которые необходимо улучшать.

В современных условиях, когда часто изменяются законодательные требования к отчетной документации в вузах, происходит непрерывная модификация инженерных образовательных программ [5] и возрастает нагрузка на профессорско-преподавательский состав [6] по разработке методических указаний по преподаваемым дисциплинам, рабочих программ, учебно-методических пособий, учебно-методических комплексов, статей для научных журналов, тезисов для конференций и т. д., становится все более актуальной проблема автоматизированного аннотирования документов.

В связи с актуальностью указанной проблемы авторы разработали алгоритмы программной реализации автоматизированного реферирования (аннотирования) документов.

Алгоритм автоматического реферирования/аннотирования

Если представить для наглядности всю схему реализованной программы в виде линейного алгоритма, то она выглядит в виде следующей последовательности действий:

- чтение текста;
- удаление из текста специальных символов;
- разбивка на слова (получение массива слов);
- для каждого слова получение леммы – преобразование слова к нормализованной форме [7, 8];
- определение частотности повторения лемм в тексте;

- выборка слов, у которых частота в тексте максимальная;

- все предложения, в которых встречается минимум три ключевых слова, попадают в аннотацию;

- вывод аннотации.

Для начала работы с системой необходимо выполнять запуск исполняемого файла разработанного приложения.

После запуска приложения отображается форма с вкладками (рис. 1):

- Входной текст/ документ;

- Ключевые слова;

- Аннотация.

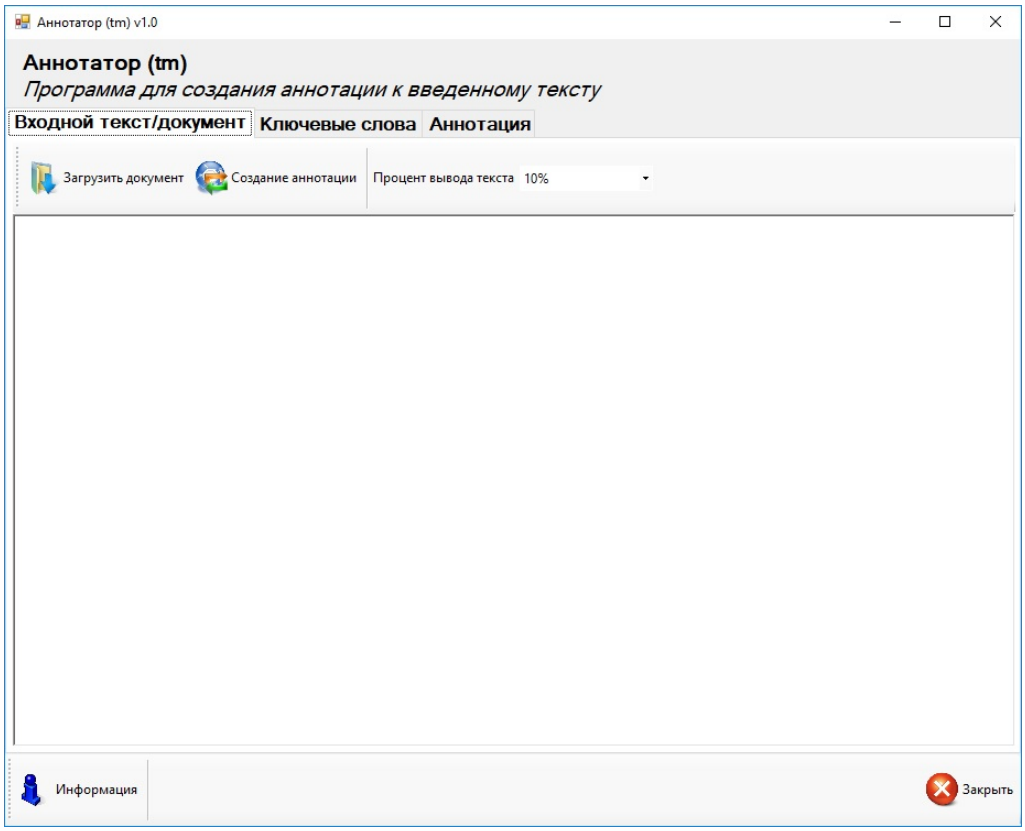

Рис. 1. – Стартовое окно

На форме присутствуют две кнопки – «Загрузить документ» и «Создание аннотации». После нажатия кнопки «Загрузить документ» появится окно выбора файла (рис. 2), где пользователь указывает необходимый ему документ.

Выбранный документ должен иметь формат «rtf». Это связано с тем, что rtf -формат обеспечивает возможность обмена текстовыми документами между различными прикладными программами Windows, так как фирма Microsoft определила этот формат в качестве Clipboard-формата. Кроме того, текстовые документы с rtf-расширением поддерживают многие типы шрифтов и атрибуты форматирования, позволяют отображать рисунки, диаграммы, таблицы, а также содержат команды управления.

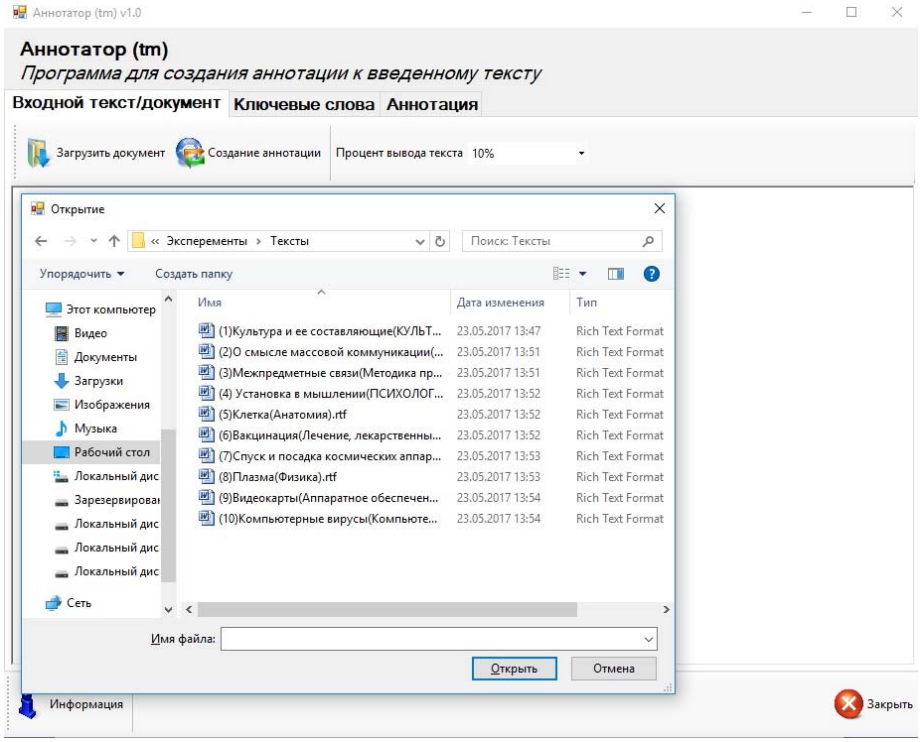

Рис. 2. – Экранная форма выбора документа

После загрузки файла текст отображается в окне, где пользователь может просматривать и редактировать загруженный текст. При нажатии на кнопку «Создание аннотации» начинается анализ текста.

Первым шагом при создании аннотации является преобразование всех слов в тексте в леммы, то есть текст вида «Письмо датировано вчерашним днем» после лемматизации будет выглядеть как «Письмо датирование вчерашний день».

Следующим шагом подсчитывается частота повторений в тексте этих лемм. Далее происходит анализ: выбираются наиболее часто повторяемые леммы и сравниваются с базой данных; если в базе данных эта лемма часто повторяется, то она просто слово-паразит, если частота в базе данных низкая, значит, слово является ключевым. База данных, по сути, нужна лишь для хранения леммизированных слов. Далее формируется список ключевых слов. Затем выбираются предложения, в которых встречаются по несколько ключевых слов. Таким образом, аннотации формируются только из тех предложений, в которых высокая концентрация слов из обсуждаемой темы. Самые первые отобранные предложения содержат наибольшее количество ключевых слов [9], то есть они содержат основной смысл статьи.

Размер аннотации зависит в процентном соотношении от исходного текста и не может быть более 3% оригинала [10] для больших текстов. Поэтому вводятся диапазоны выводимой аннотации: 3%, 5% [11] или 10% от размера исходного файла.

В программе выбираются предложения, в которых встречаются ключевые слова, далее в зависимости от процента выбранного вывода отбирается 3, 5 или 10 процентов этих предложений. В программе реализовано выпадающее меню с выбором процента вывода размера аннотации от исходного размера документа (рис. 3).

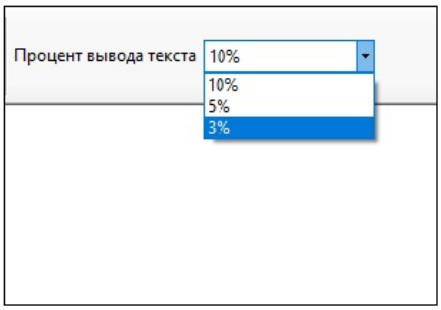

Рис. 3. – Экранная форма выбора процента вывода текста

Когда размер документа очень большой и количество страниц в документе настолько много, что даже 3 процента от исходного текста – это

не одна и не две страницы аннотации. Такая аннотация превращается в реферат. Данная проблема была решена ограничением получаемого текста.

Установлена верхняя граница выводимой аннотации: 10 предложений – максимальный размер получаемой аннотации. Текст длиной в 2000 символов назвать аннотацией уже нельзя. Поэтому было реализовано ограничение сверху для сохранения смысла разрабатываемой программной реализации.

После успешно выполненного анализа документа, во вкладке «Ключевые слова» можно просмотреть весь список сгенерированных ключевых слов (рис.4), состоящий из трёх и более слов, наиболее релевантных, поскольку, по рекомендациям, требуется от трех до пяти ключевых слов.

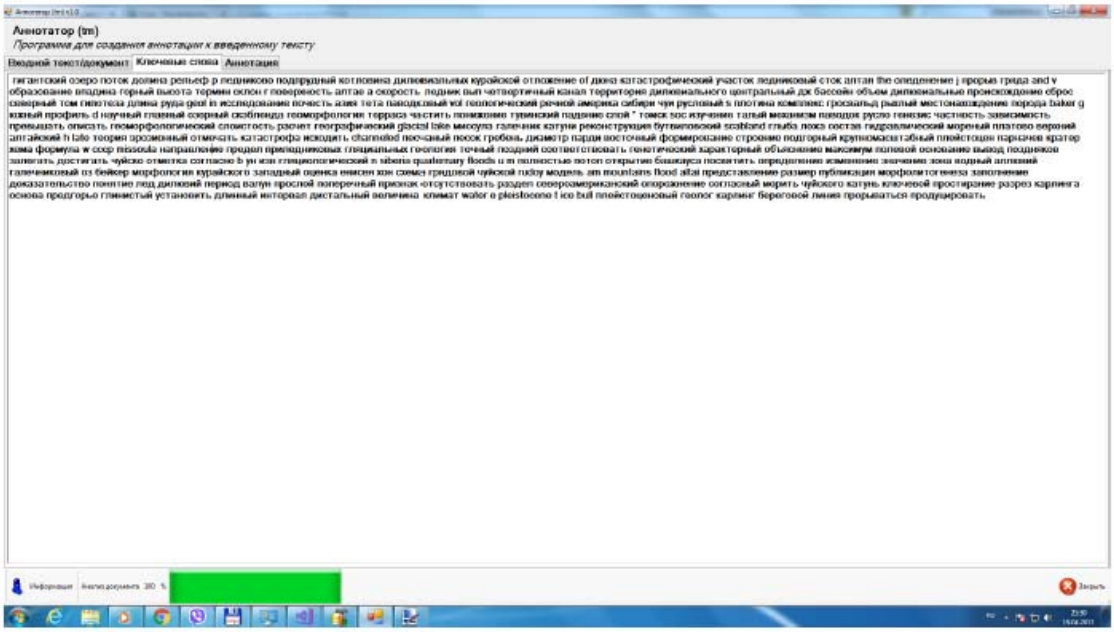

Рис. 4. – Выявленные ключевые слова в тексте документа

В последней вкладке «Аннотация» (рис.5) можно просмотреть полученную аннотацию по заданным параметрам и исходным данным.

Также реализована кнопка «Закрыть» (рис. 6), при нажатии на которую происходит выход из программы «Аннотатор» без сохранения результатов работы, а затем появляется диалоговое окно (рис. 7) с запросом о закрытии приложения.

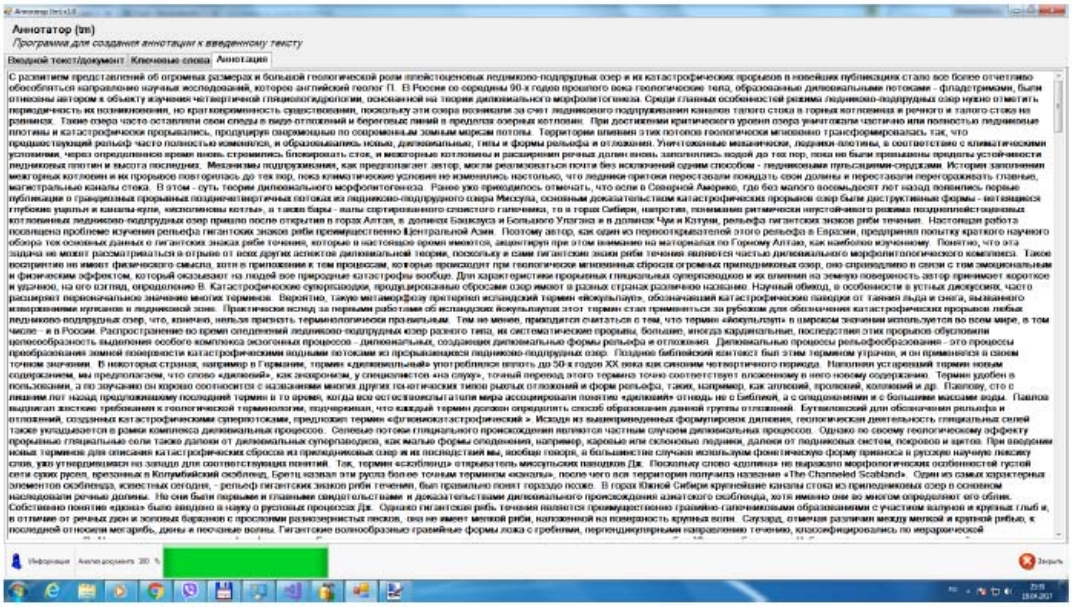

Рис. 5. – Экранная форма вывода аннотации

Рис. 6. – Выход из программы

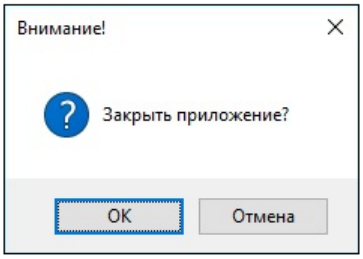

Рис. 7. – Закрытие приложения

Для оценки эффективности реализованной автоматизированной системы аннотирования документов использовался сервис Главред в качестве автоматизированного эксперта. Оценка по шкале Главреда — соотношение стоп-слов – слов, не несущих смысловой нагрузки, к остальным словам. Чем выше оценка, тем меньше в тексте стоп-слов.

С помощью разработанного программного средства составлены 10 аннотаций для технических текстов и 10 аннотаций для документов, относящихся к художественной литературе. На рис. 8 изображён график с каждой итерацией эксперимента для технической и художественной литературы. Среднее значение, полученное при оценке технического текста

равно 8,08 балла, а при оценке художественного текста – 8,06 из 10 возможных.

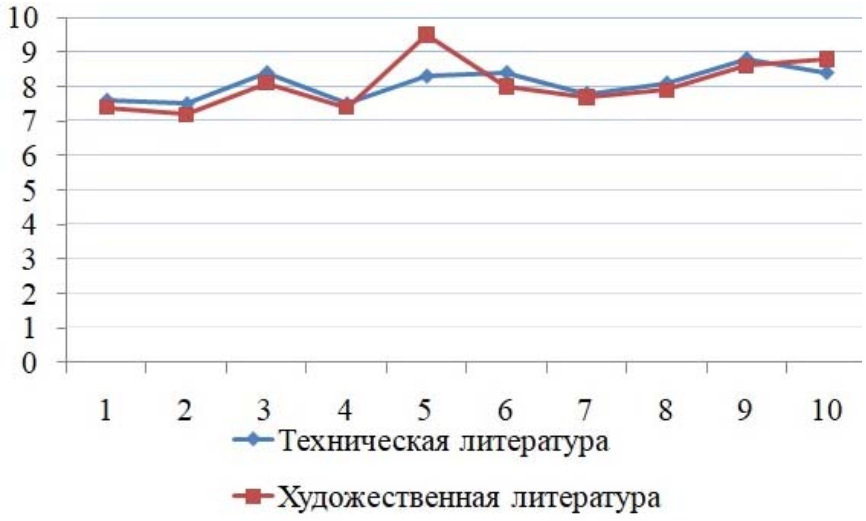

Рис. 8. – Результаты эксперимента

Разница составляет 0,02. Это говорит о том, что качество полученных аннотаций не зависит от вида исходной литературы.

Заключение

Представленный тестовый программный продукт для автоматизированного аннотирования документов позволяет выполнять предварительную обработку текстового документа, анализ текста документа, отбор фрагментов текста, содержащих релевантную для аннотации информацию, а также вывод ключевых слов и генерацию текста аннотации. Использование данного программного средства обеспечивает возможность существенно сократить время на оформление и подготовку к печати текстовых документов.

Таким образом, современные системы автоматизированного аннотирования способны оказать важную услугу людям, чья профессиональная деятельность связана с анализом большого количества информации. У данного научно–инженерного направления имеется большое количество перспективных путей развития.

Литература

1. Тарасов С.Д. Современные методы автоматического реферирования // Системный анализ и управление . 2016. №1. С. 59-74.

2. Borko H., Bernier Ch. Abstracting concepts and methods. New York: 2007. 250 p.

3. Rybanov A.A., Makushkina L.A. Technology of an aprioristic objective assessment of distance course themes complexity based on Saati's algorithm // Journal of Engineering Science and Technology Review . 2016. №1. pp. 81-89.

4. Бурмистров А.С., Свиридова О.В. Экспертная оценка программных продуктов для аннотирования документов // Постулат. 2017. № 5 URL: e-postulat.ru/index.php/Postulat/article/view/567/588.

5. Моисеенко Н.А. Трансформационное обучение и холистический подход в информационно-образовательной среде технического вуза // Инженерный вестник Дона. 2013. №4 URL: ivdon.ru/ru/magazine/archive/n4y2013/2034.

6. Анисимова Г.Б., Романенко М.В. ИС автоматизации формирования учебно-методических материалов в условиях реформы Высшей школы // Инженерный вестник Дона. 2013. №4 URL: ivdon.ru/ru/magazine/archive/n4y2013/2147.

7. Осминин П.Г. Построение модели реферирования и аннотирования научно-технических текстов, ориентированной на автоматический перевод: автореф. дис. ... канд. филол. наук: 10.02.21. Челябинск, 2016. 24 с.

8. Лемматизация // Лемматизация в API грамматического словаря URL: solarix.ru/for_developers/api/lemmatization.shtml (дата обращения: 29.05.2018).

9. Мащенко Н.Г. Исследование и разработка системы автоматического реферирования текстов на основе ранжирования связных структур // Донецк: ДонНТУ, 2013 URL: masters.donntu.org/2013/fknt/mashchenko/diss/index.htm.

10. Орлова Е.А. Научный текст: аннотирование, реферирование, рецензирование. Учебное пособие для иностранных студентов-медиков и аспирантов. СПб.: Златоуст, 2015. 100 с.

11. Ефремова М. И. Автоматический разбор и аннотирование статей // Фундаментальные исследования. 2015. №2 URL: fundamental-research.ru/pdf/2015/2-22/38121.pdf .

References

1. Tarasov S.D. Sistemnyj analiz i upravlenie [System analysis and management]. 2016. №1. pp. 59-74.

2. Borko H., Bernier Ch. Abstracting concepts and methods. New York: 2007. 250 p.

3. Rybanov A.A., Makushkina L.A. Journal of Engineering Science and Technology Review. 2016. №1. pp. 81-89.

4. Burmistrov A.S., Sviridova O.V. Postulat. 2017. № 5 URL: e-postulat.ru/index.php/Postulat/article/view/567/588.

5. Moiseenko N.A. Inženernyj vestnik Dona (Rus). 2013, №4 URL: ivdon.ru/ru/magazine/archive/n4y2013/2034.

6. Anisimova G.B., Romanenko M.V. Inženernyj vestnik Dona (Rus). 2013. №4 URL: ivdon.ru/ru/magazine/archive/n4y2013/2147.

7. Osminin P.G. avtoref. dis. ... kand. filol. nauk: 10.02.21. - Cheljabinsk, 2016.

8. Lemmatizacija [Lemmatization]. Lemmatizacija v API grammaticheskogo slovarja [Lemmatization in the grammar dictionary API]. URL: solarix.ru/for_developers/api/lemmatization.shtml.

9. Mashhenko N.G. Doneck: DonNTU, 2013 URL: masters.donntu.org/2013/fknt/mashchenko/diss/index.htm.

10. Orlova E.A. Nauchnyj tekst: annotirovanie, referirovanie, recenzirovanie. Uchebnoe posobie dlja inostrannyh studentov-medikov i

aspirantov. [Scientific text: annotation, abstracting, reviewing. Textbook for foreign medical students and postgraduates]. SPb.: Zlatoust, 2015. 100 p.

11. Efremova M. I. Fundamental'nye issledovanija. 2015. №2. URL: fundamental-research.ru/pdf/2015/2-22/38121.pdf.