DOI: 10.17117/na.2016.04.03.205

Поступила (Received): 24.04.2016

<http://ucom.ru/doc/na.2016.04.03.205.pdf>**Суворова А.В.****К вопросу об автоматизации формирования
гlossариев узкоспециализированных терминов****Suvorova A.V.****The way of automating the creation of a glossary which
consists of terms from particular domain of knowledge**

В данной статье предложено решение актуальной проблемы – повышения качества перевода научных статей на иностранный язык. На основе проведенного исследования автором предлагается решение, основанное на автоматизированном составлении glossария терминов по узкой теме.

Использование узкоспециализированных glossариев поможет правильно использовать термины

Ключевые слова: алгоритм, glossария, обработка данных, набор данных

This article provides a solution to the urgent problem. It shows how to improve the quality of scientific article's translation. On the basis of the research the author suggests using a special method based on the automated creation of a glossary which consists of terms from particular domain of knowledge. We will choose correct term, if we use a special glossary

Key words: algorithm, glossary, data processing, dataset

Суворова Арина Валерьевна

Студент

Санкт-Петербургский национальный
исследовательский университет информационных
технологий, механики и оптики

г. Санкт-Петербург, Кронверкский пр., 49

Suvorova Arina Valerievna

Student

Saint-Petersburg national research university of
information technologies, mechanics and optics
St. Petersburg, Kronverksky ave., 49

В эпоху всемирной глобализации студентам и преподавателям российских университетов необходимо стать активными участниками международного научного сообщества. Все это немыслимо без взаимодействия с другими университетами, изучения зарубежных источников информации, публикаций в зарубежных изданиях и участия в международных конференциях.

Следует отметить, что российские университеты успешно интегрируются в мировое образовательное и научное пространство: происходит обмен студентами с зарубежными вузами, приглашаются лучшие профессора общепризнанных мировых университетов, наши студенты принимают участие в различных международных межвузовских олимпиадах. Однако с публикациями в зарубежных изданиях дело обстоит сложнее. Интернациональным языком науки по общему признанию является английский. Вполне естественно, что США лидируют по числу написанных и опубликованных в научных журналах статей (536 230 публикаций за 2015 год), а наша страна занимает лишь 15-е место в общемиро-

вом списке (общее количество публикаций российских ученых за этот же год составило 50 208). Если проанализировать информацию из базы данных Scopus, в которой общее количество публикаций составляет 2 363 947, то получаем, что доля российских статей равна всего 2,12%.

Анализ исследований, посвященных проблеме публикации научных статей в зарубежных изданиях, и различных источников, в которых указываются необходимые требования для публикации, позволил выявить основные причины отказов:

- несоблюдение принятых стандартов построения и оформления текстов;
- публикация может быть неинтересна журналу, так как не соответствует его профилю;
- несоответствие мировым стандартам и критериям научности;
- низкое качество статьи, или она содержит плагиат;
- плохой перевод на английский язык, а именно то, что узкоспециализированная лексика переведена неправильно.

Работа посвящена решению последней из перечисленных проблем. Решением является автоматизированное формирование глоссария узкоспециализированных терминов на основе изучения научных текстов, представленных в международных базах, посвященных конкретной тематике исследования.

Следует отметить, что понятие «глоссарий» имеет несколько определений. В рамках работы мы будем понимать под глоссарием толковый словарь терминов по определенной узкой тематике с переводом на иностранный язык.

Решение проблемы автоматизации формирования глоссария мы видим в следующем алгоритме:

- определение тем, по которым будет составлен набор данных;
- выбор ресурса, с помощью которого будет составлен набор данных;
- формирование набора данных на основе извлечения статей, которые удовлетворяют нашим интересам, с помощью парсера;
- предобработка статей;
- применение алгоритма извлечения терминов к найденным статьям;
- фильтрация полученных списков ключевых выражений;
- добавление оставшихся терминов в глоссарий;
- проверка полученного глоссария.

Процесс выделения терминов и составления на их основе глоссария состоит из нескольких этапов, которые представлены на рисунке 1.1.

Кратко опишем каждый из этапов:

1. Получаем набор данных, включающий в себя статьи по определенной узкой тематике.

2. Предварительная обработка текста. На этом этапе происходит подготовка текста к дальнейшим преобразованиям, а именно: удаляются стоп-слова, не несущие смысловой нагрузки (местоимения, предлоги, частицы, союзы, междометия и т.д.), слова приводятся к начальной форме.

Рис. 1.1. Процесс выделения терминов из текста

3. Отбор кандидатов в ключевые слова зависит от метода извлечения ключевых слов. Если мы используем статистический метод, то кандидаты отбираются на основании данных о частоте встречаемости слова или выражения в тексте. В случае, когда мы используем лингвистический метод, применяются онтологии и семантические данные о слове.

4. Ранжирование кандидатов по частоте встречаемости. Получаем список, в котором ключевые слова расположены от наиболее часто до наиболее редко встречающихся выражений.

5. Пост-обработка. Не все найденные выражения будут являться ключевыми. Поэтому необходимо провести ручную фильтрацию ложных ответов.

6. Представление результатов. По каждому тексту имеем список его ключевых терминов, которые включаем в узкоспециализированный глоссарий.

В работе представлен обобщенный алгоритм формирования глоссария узкоспециализированных терминов, но в большинстве случаев этапы зависят от

выбранного метода преобразования. Таким образом, может быть изменена последовательность, убраны или добавлены некоторые пункты.

© 2016, Суворова А.В.

*К вопросу об автоматизации формирования
гlossариев узкоспециализированных терминов*

© 2016, Suvorova A.V.

*The way of automating the creation of a glossary
which consists of terms from particular domain of
knowledge*