УДК 81’33 + 811.811.11:004.4

СЕГМЕНТАЦИЯ И ТОКЕНИЗАЦИЯ ТЕКСТА НА АРАБСКОМ ЯЗЫКЕ С ПРИМЕНЕНИЕМ ПРОГРАММНОГО КОМПЛЕКСА NLTK

А.Р. Гареева

В настоящей статье описаны модули программного комплекса NLTK, приведены результаты апробации таких модулей как токенизация и сегментация текста на арабском языке.

Ключевые слова: NLP, обработка естественного языка, NLTK, модули Python, арабский язык, токенизация текста, сегментация текста.

Как известно, прикладная лингвистика помимо направлений, связанных с теорией и практикой преподавания языков, включает в себя такое направление, как автоматическая обработка естественного языка (Natural Language Processing, NLP) – общее направление искусственного интеллекта и компьютерной лингвистики.

На сегодняшний день большая часть исследований российской лингвистики связана с западными языками. Исследования неевропейских языков встречаются достаточно редко. Особенно это касается исследований в области компьютерной лингвистки, а именно разработок программного обеспечения по обработке текстов на восточных языках.

В своей статье Фаткулин Б. Г. обосновал необходимость использования достижений прикладной лингвистики для эффективного сбора информации на восточных языках [1].

В настоящее время существуют многочисленные программные средства для автоматической обработки текстов. Эти средства предназначены для широкого круга задач: машинный перевод, автоматическое реферирование, создание баз знаний, разметка корпусов текстов, поиск информации и др. Важное место занимают расширяемые системы, предназ-

наченные для обработки, задаваемой и настраиваемой пользователем. Наиболее популярные из них – это GATE, UIMA, NLTK [1].

При исследовании больших текстовых массивов на восточных языках, наше внимание было обращено на программный комплекс NLTK (Natural Language Toolkit) – пакет библиотек и программ для символьной и статической обработки естественного языка, которые были написаны на языке программирования Python. NLTK был разработан в 2001 году в Пенсильванском исследовательском университете командой прикладных лингвистов (Steven Bird, Ewan Klein, Edward Loper, Jason Baldridge) как составная часть учебного курса по компьютерной лингвистике. В настоящее время NLTK используется в качестве учебного пособия в десятках высших учебных заведениях и служит основой многих научно-исследовательских проектов.

Целью нашего исследования стала апробация NLTK в применении к языкам, использующие нетрадиционные формы представления текстов, в нашем случае для арабского языка.

- изучить модули NLTK на языке программирования Python;
- апробировать исследуемые модули NLTK в применении к арабскому языку.

Natural Language Processing (обработка естественного языка, NLP) включает в себя множество задач, таких как морфологический и синтаксический анализы, извлечение ключевых слов, извлечение именованных сущностей (Named Entity Recognition), построение конкордансов, автоматическое реферирование (Automatic Text Summarization), токенизация (tokenization) и т.д. На сегодняшний день существуют инструменты, с помощью которых эти задачи успешно решаются. Используемый нами комплекс программ NLTK имеет различные инструменты (модули) для символьной и статической обработки текстов.

Пакет библиотек и программ NLTK создан для работы в основном с индоевропейскими языками, в частности для обработки текстов на английском языке [2].

В нашей исследовательской работе мы использовали некоторые модули NLTK для работы с восточными языками, имеющих нетрадиционные средства представления информации, а именно арабскую вязь.

В качестве учебного пособия использовалась книга *Natural Language Processing with Python,* написанная командой американских лингвистов, которые создали комплекс программных модулей NLTK [3].

В настоящей статье рассматриваются два инструмента: токенизация и сегментация текста.

Токенизация – разбиение потока символов в естественном языке на отдельные значимые единицы (токены, словоформы) [4]. Сложность данной процедуры заключается в том, что языки не обладают совершенной пунк-

туацией. Это вызывает определенные трудности у программы, поскольку из-за наличия в текстах сокращений, аббревиатур и т.д. некоторые случаи не могут быть однозначно токенизированы.

Сегментация в лингвистике – линейное членение речевого потока на составляющие отрезки [5], называемые сегментами.

Для проведения процедуры токенизации необходимо предварительно «обучить» программу определённым командам.

В качестве материала для работы был использован online-текст № 43007 «Tribute to Michael Hart, by Majid AlHydar» с электронной библиотеки проект «Гутенберг».

При изучении программного комплекса было выявлено два способа токенизации текста:

1. В скрипт вставляется ссылка на онлайн-текст.
2. В скрипт вставляется необходимый для токенизации текст.

В качестве примера приведем полученный нами первым способом токенизированный текст.

Для проведения процедуры токенизации необходимо предварительно «обучить» программу определённым командам.

```
>>> from nltk import word_tokenize
>>> url = "http://www.gutenberg.org/files/43007/43007-0.txt"
>>> raw = response.read().decode('utf8')
>>> tokens = word_tokenize(raw)
>>> tokens[100:111]
['أحدث', 'ثـ ورة', 'في', 'مـ شروع', 'هلرت', 'الـ ذي', 'نـ ما', 'بـ عدها', 'بـ سرعة', 'عظـ يمة',
'حـ تى',]
```

Следующий этап работы с текстом – его сегментация. В нашей исследовательской работе было разобрано и описано несколько подходов к сегментации текста.

1. Сегментация текста на предложения. Для данной процедуры NLTK предлагает использовать встроенный в программу Punkt sentence segmenter [6].

2. Сегментация текста на словосочетания и слова. Такой способ сегментации подразумевает предварительную подготовку программы с помощью ручного ввода данных.

Прежде чем приступить к работе в сегментере, необходим ввод специального кода. Пример кода имеется в открытом доступе на сайте NLTK [3], и его можно скачать, кликнув на ссылку под разделом «Segmentation».

```
>>> raw = """هذا أخ بار تـ تـ لـ قف الـ تي الإعـ لام أ ضواء عن بعيداً ..وهدوء صمت فـ ي
والـ سد يـ نما والـ سد يا سة الـ ريـ ا ضة نـ جوم من وذاك..."""
>>> text = raw
>>> sents = nltk.sent_tokenize(text)
>>> pprint.pprint(sents)
```

['\n'
' ذ جوم من وذاك هذا أخ بار ت ت ل قف ال تي الإع لام أ ضواء عن بعيداً ..وهوء صمت ف ي'

Это один из самых простых способов сегментации: в учебном пособии представлены множества кодов и вариантов их применения для решения разного рода лингвистических задач.

Таким образом, подводя итог нашего исследования, мы пришли к выводу, что использование модулей NLTK позволяет обрабатывать информацию на арабском языке. Однако данное программное обеспечение требует дальнейшей доработки, с учетом всех языковых особенностей арабского языка. Тем не менее, на сегодняшний день существует множество лингвистически ориентированных программных модулей и инструментов, направленных на работу с информацией на восточных языках, и учитывающих все их морфологические и синтаксические особенности.

Библиографический список

1. Фаткулин, Б.Г. Прикладная лингвистика и обработка текстов на восточных языках: современные перспективы / Б.Г. Фаткулин // Вестник ЮУрГУ. Серия «Лингвистика». – 2014. – Вып. 11. – № 3. – С. 15–18.

2. Арабский язык. Энциклопедия «Кругосвет» [Электронный ресурс]. – URL: http://www.krugosvet.ru/node/31404?page=0,2/.

3. Bird, S. Natural language processing with Python / S. Bird, E. Klein, and E. Loper. – Beijing; Cambridge; Mass: O'Reilly, 2009. Print. Garrette, D. An extensible toolkit for computational semantics / D. Garrette, E. Klein // Proceedings of the Eighth International Conference on Computational Semantics (IWCS-8 '09) / H. Bunt, V. Petukhova, S. Wubben (Eds.). Association for Computational Linguistics, Stroudsburg, PA, USA, – P. 116–127.

4. Finegan, E. LANGUAGE: its structure and use / E. Finegan. – New York: Harcourt Brace College Publishers, 2004. – 613 p.

5. Сегментация. (Лингвистика) – Википедия. Лингвистический энциклопедический словарь. [Электронный ресурс]. – URL: https://ru.wikipedia.org/wiki/.

6. Kiss, T. Unsupervised multilingual sentence boundary detection / T. Kiss, J. Strunk // Comput. Linguist. – 2006. – Pp. 485–525.

К содержанию