Компьютерная паралингвистика текстов контактных языков

В. С. Крылов*

Крымский инженерно-педагогический университет, Симферополь, Крым
*Corresponding author. E-mail: vladimi-krylov@yandex.ru

Paper received 25.07.15; Accepted for publication 04.08.15.

Аннотация. Информационные технологии (ИТ) сформировали глобальную инфраструктуру связи между людьми. В сетях интернета возникли и активно развиваются контактные языки. Синтаксис и грамматика таких языков отличается от синтаксиса и грамматики базового языка. Она подобные пиджин и креольским языкам. Язык программирования Python с библиотекой автоматизированного анализа естественного языка NLTK предоставляет обширный набор инструментов для автоматического анализа контактных языков. Эффективный анализатор структуры текстов таких языков - обработчик событий.

Ключевые слова: *компьютерная паралингвистика, событийно-ориентированный подход, креализованный язык*

Информационные технологии (ИТ) сформировали глобальную инфраструктуру, обеспечивающую передовые услуги за счет организации связи между людьми, и даже между вещами физическими или виртуальными. Общение в сети интернет, обмен текстами в сети повлёк за собой появление новых языковых особенностей, лингвистические характеристики которых определяются соответствующим каналом коммуникации в информационном пространстве. Такими, как электронная почта, мгновенный обмен сообщениями в чатах, СМС сообщения, твитер и так далее. Просто текстовых сообщений для передачи, например, эмоционального состояния коммуниканта оказалось недостаточно. В результате, спонтанно возникли смайлики. Текст сообщения стал дополняться кодами, не имеющими прямого отношения к алфавиту и грамматике сообщения. Синтаксис, грамматика таких сообщений значительно отличаются от синтаксиса, грамматики базового языка коммуникантов. Эти языки определяются как контактные креализованные языки, подобные пиджин и креольским языкам [6, 7].

Совершенствование информационных технологий (ИТ) предоставило широкому кругу исследователей доступные аппаратные и программные средства сбора, обработки исходных данных и их анализа. Появилась возможность использовать ранее недоступные инструменты и технологии для экспериментальных и теоретических исследований в самых различных научных направлениях. В первую очередь, в междисциплинарных исследованиях. На основе этих технологий возникают новые интегрированные научные направления. Например, компьютерная паралингвистика предлагает модели, которые позволяют значительно усовершенствовать системы анализа и обработки естественной речи и текстов, обработки сообщений в социальных сетях, усовершенствовать поисковые системы [3, 13].

Цель статьи – предложить модель организации и анализа сообщений креализованных текстов на основе моделей компьютерной паралингвистики и объектно-ориентированного и событийно-ориентированных подходов в программировании.

Язык, как система фонетических, лексических и грамматических средств, порождает естественную информационную реальность, через которую обеспечиваются коммуникации в обществе. Внедрения информационных технологий (ИТ) в самые различные сферы деятельности людей, стремительное развитие сети интернет первоначально рассматривалось как дополнение или расширение естественного информационного пространства. Совершенствование технических и программных средств из дополнения естественного информационного пространства превратилось в самостоятельную информационную реальность, а общество незаметно для себя стало информационным.

В каналах мгновенных текстовых сообщений интернет чатов, твиттере, форумов, СМС и других идет активное словообразование, например, через образования новых аббревиатур и сокращений. В тексты обменов активно включаются смайлики, визуальные и аудио элементы. Текст сообщения становится поликодовым, в котором все элементы не являются просто суммой, объединением, а интегрированы единое целое для передачи смысла. Такие тексты определяются как креализованные [6]. То есть тексты, состоящие из двух объединённых в единую структуру частей: вербальной (речевой) и невербальной, которая относится к другим знаковым системам. Обе эти части зависят друг от друга. Вербальную часть невозможно отделить от невербальной [6, 12, 13].

Рассмотрим, например, обмены сообщениями в популярной социальной сети Инстаграм. Она предназначена для размещения собственных фотографий о каких-либо событиях, пейзажах и личных изображений на этом фоне [1]. В конечном итоге фото размещают для получения отзывов или комментариев, как от всех возможных участников сети, так и для определенного круга пользователей. Комментарии являются откликами, и вполне соответствуют определению понятия события, и составляют поток событий. Они сами по себе могут вызывать отклики, то есть являются событием, которое представляет определенную реакцию на данный комментарий [2, 3].

Событие, как понятие, имеет различные определения в различных отраслях деятельности, ситуациях, обстоятельствах. В данном случае событие представляет собой сообщение, посылаемое объектом, чтобы сигнализировать о совершении какого-либо действия. Любой объект, способный вызывать события, является *отправителем событий*, также называемым *источником событий*. Отправителю события зачастую не известен объект, который будет получать, и обрабатывать сформированные отправителем события. Получатель должен обладать о*бработчиком событий* – процедурой, исполняемой при получении сообщения о соответствующем событии [2, 9].

Поток комментариев представляет собой поток событий. Во всяком случае, участники сети воспринимают их как события, и принимают решение как на них реагировать. Ответ или отклик определяется в

соответствии с представлениями об объекте и внутренним состоянием на момент ответа. Он может быть представлен текстом, фото, видео, смайликом, а также какой-либо их комбинацией. Сообщение комментария представляет собой некоторый креализованный текст, отображающий вербальную и невербальную оценку события. Каждая из этих компонент формирует сообщение по своим собственным правилам. Синтаксис и грамматика вербальной части, как правило, отличаются от синтаксиса и грамматики базового языка. Существует определенная зависимость между получаемым, например, визуальным событием и структурой предложения отклика на событие [10].

Размещенные в текстах комментариев, сообщений различные знаки, символы и смайлики имеют определенные коды, поэтому возможно составлять единые шаблоны поиска и анализа сообщений, составленных разными знаковыми системами. То есть возможен последовательный поиск, анализ собственно текстов и отдельно последовательностей символов или знаков, отображающих невербальную компоненту сообщения. И, кроме того, возможен поиск и анализ организации объединенной вербально-невербальной последовательности символов [3].

Поиск в текстах и анализ текстов по единому шаблону, объединяющему вербальные и невербальные компоненты, лежит в основе восприятия текстов людьми. Результаты исследований с помощью компьютерной томографии процессов заучивания новых слов показали, что каждое слово из незнакомой последовательности букв постепенно превращаются в единый визуальный образ [5]. В начале новое слово воспринимается как нечто неизвестное, состоящее из отдельных знаков, но потом оно становится знакомым, и входит в словарь единым визуальным изображением. Затем происходит переход от чтения по буквам к восприятию слова как единого визуального образа. По сути дела, в процессе изучения слов формируется «визуальный словарь». Слова распознаются сразу, как единое целое и целиком в участке левой зрительной коры. Этот участок симметричен участку зрительной коры справа, отвечающему за распознавание лиц. Зрительные нейроны постепенно настраиваются на новое видение текста, происходит переход от чтения по буквам к распознаванию слова как единого целого. Очевидно, что это ускоряет и облегчает чтение, как и узнавание лиц не по отдельным чертам, а целиком, помогает быстро узнавать другого человека и облегчает общение [5]. Таким образом, модель представления структуры сообщений на креализованном языке, как последовательности некоторых шаблонов, объединяющих разные знаковые системы, отображает реальные процессы распознавания человеком слов, предложений и текстов.

В разработке систем распознавания текстов в настоящее исходят из представления о побуквенном чтении текстов, побуквенном разборе слов в предложениях. Модель структуры сообщений на креализованном, как последовательности некоторых шаблонов, объединяющих разные знаковые системы, дает возможность в едином ключе, на единой основе обеспечивать интегрированное целостное представление сообщения, объединить лингвистические паралингвистические факторы вербального поведения. Очевидно, что для выделения объединений из разных знаковых систем, шаблоны могут быть разного уровня сложности. Они являются источником моделей анализа структуры реальных сообщений на контактных языках. Например, для упрощенного креализованного языка с некоторым ограниченным набором слов и смайликов или некоторых изображений. Такие языки распространены в рекламной деятельности [6].

Для поиска и анализа синтаксической структуры самых разнообразных текстов, размещенных в сети интернет активно используются программы парсеры. Эти программы разрабатываются для автоматизированного поиска, сбора и анализа различной информации: текстов кодов программ, фильмов, музыки, тематических архивов документов, отдельных документов, рецензий на события и так далее. Процесс поиска и анализа такой программой определяется как парсинг. Поэтому в основном под парсингом обычно понимается визуальный, автоматический синтаксический и лексический анализ текстов, разбор какого-либо документа для извлечения из него определенных данных. Часто востребованный результат парсинга заключается в поиске контента, содержащего не только требуемые ключевые слова, но и определенную синтаксическую организацию текста.

Особенность исполнения парсинга определяется заданием схемы отбора информации. Параметры поиска и анализа задаются заранее. Собранную информацию парсер предоставляет в определенном виде. Например, результатом работы парсера может быть база данных [11].

Наиболее часто парсинг опирается на анализ по образцам, составленным из регулярных выражений. Регулярные выражения - система обработки текста, основанная на специальной системе записи образцов для поиска [9, 11]. Образец или паттерн, задающий правило поиска, называют шаблоном или маской. Регулярные выражения используются многими текстовыми редакторами и утилитами для поиска и изменения текста на основе выбранных правил. В настоящее время сложно найти язык программирования, который не поддерживает регулярные выражения для работы со строками, составленные из различных символов. Язык, который не имеет встроенный синтаксис и механизм обработки регулярных выражений. Регулярные выражения используются для сжатого описания некоторого множества строк с помощью шаблонов, без необходимости перечисления всех элементов этого множества, и при составлении, которых используется специальный синтаксис.

Шаблоны из регулярных выражений образуют некоторый мета язык. Собственно, этот мета язык и образует модель организации анализируемых сообщений, представленных некоторым креолизированным языком. Кроме того, такой шаблон означает некоторую структуру, которая воспринимается как событие. Анализ последовательностей встретившихся единиц сообщения, соответствующих некоторому шаблону, требует обработчика событий. Поэтому анализ структуры сообщений креализованного текста будет являться результатом работы этого обработчика событий.

Язык программирования Python с открытой библиотекой автоматизированного анализа естественного языка NLTK предоставляет обширный набор инструментов для работы с регулярными выражениями [8].

Библиотека NLTK является практически идеальной платформой разработки парсеров поиска контента, содержащего не только ключевые слова, определенную синтаксическую организацию вербальной компоненты текста, но и его невербальный компонент. Открытый код модулей, их подробное документирование предоставляют исключительные возможности в реализации интегрированных моделей анализа креализованных языков коммуникаций. Эти модели имеют научное и прикладное значение для разработки новых и совершенствования существующих информационных сис-тем.

На платформе Python + NLTK можно решать различной сложности задания из области компьютерной паралингвистики, в том числе и разработанные для учебного процесса подготовки специалистов в области информационных технологий. Выполнение таких заданий стимулирует развитие навыков и умений приобретать дополнительные знания, осваивать как можно больше востребованных информационных технологий, умение разрабатывать программные продукты для разных отраслей, сфер и направлений практического их применения. Дает возможность будущему специалисту приобрести навыки эффективного решения поставленных заказчиком задач. Именно на таком материале, на основе приобретенного опыта, молодой специалист получает необходимый набор профессиональных навыков в успешной конкуренции на ИТ-рынке труда [4].

ЛИТЕРАТУРА

[1] Все о программе Инстаграм (Instagram) по-русски / URL: http://instagrama.ru/

[2] онцептуальная модель для систем обработки событий / URL: http://www.ibm.com/developerworks/ru/library/ws-eventprocessing/index.html

[3] Крылов В.С. Компьютерная паралингвистика как основа приобретения специальных профессиональных навыков в области информационных технологий / Крылов В.С. – Science and Education a New Dimension. Philology, II(4), Issue: 24, 2014 С. 41-44.

[4] Крылов В.С. Молодые специалисты способны конкурировать на рынке труда информационных технологий. / Крылов В.С. – Science and Education a New Dimension: Pedagogy and Psychology. Vol. 5, 2013 – P. 98-101

[5] After learning new words, brain sees them as pictures / Georgetown University Medical Center (GUMC) / URL: https://gumc.georgetown.edu/news/After-Learning-New-Words-Brain-Sees-Them-as-Pictures

[6] Beisenova, Z.S.The Study of Creolized Texts in Written Communication / Z.S. Beisenova, A.T. Bayekeyeva, S.M. Sapina, B.B. Dinayeva, A.K. Utanova. – INDIAN JOURNAL OF APPLIED RESEARCH, Vol. 4, Issue 5, May 2014, p. 217

[7] Bickerton, D. Creole Languages / Bickerton D. – Scientific American, July 1983, Vol. 249, No.1, pp. 116-122

[8] Bird, S. Natural Language Processing with Python – Analyzing Text with the Natural Language Toolkit // S. Bird, E. Klein, E. Loper / URL: http://www.nltk.org/book/

[9] Event-Based Programming. Taking Events to the Limit / URL: http://sharebookfree.com/event-based-programming-taking-events-to-the-limit/

[10] Myachykov, A. Visual Cues to Structural Choice in Visually Situated Sentence Production / A. Myachykov, D. Thompson, S. Garrod, C. Scheepers. – Frontiers in Psychology, Published online Jan 18, 2012 / URL: http://journal.frontiersin.org/Journal/10.3389/fpsyg.2011.00396/full9.

[11] Kallmeyer, L. Parsing Beyond Context-Free Grammars / Kallmeyer L. – Springer-Verlag Berlin Heidelberg 2010, – 247 p.

[12] Krämer, N.C. Nonverbal Communication / N.C. Krämer. – Human Behavior in Military Contexts, P. 150-188, / URL: http://books.nap.edu/catalog/12023.html

[13] Schuller, B., Batliner, A. Computational paralinguistics : emotion, affect and personality in speech and language processing / BjoËrn Schuller, Anton Batliner. – First Edition, John Wiley & Sons, Ltd , 2014, P. 321.

REFERENCES

[1] All of the program Instagram (Instagram) in Russian / URL: http://instagrama.ru/

[2] Conceptual Model for event processing systems / URL: http://www.ibm.com/developerworks/ru/library/ws-eventprocessing/index.html

[[3] Krylov, V.S. Computer paralinguistics as the basis for the acquisition of specialized skills in information technology / V.S. Krylov – Science and Education a New Dimension. Philology, II (4), Issue: 24, 2014, P. 41-44

[4] Krylov, V.S. Young professionals are able to compete in the labor market of information technologies. / V.S. Krylov – Science and Education a New Dimension: Pedagogy and Psychology. Vol. 5, 2013, P. 98-101

[5] After learning new words, brain sees them as pictures / Georgetown University Medical Center (GUMC) / URL: https://gumc.georgetown.edu/news/After-Learning-New-Words-Brain-Sees-Them-as-Pictures

[6] Beisenova, Z.S.The Study of Creolized Texts in Written Communication / Z.S. Beisenova, A.T. Bayekeyeva, S.M. Sapina, B.B. Dinayeva, A.K. Utanova. – INDIAN JOURNAL OF APPLIED RESEARCH, Vol. 4, Issue 5, May 2014, p. 217

[7] Bickerton, D. Creole Languages / Bickerton D. – Scientific American, July 1983, Vol. 249, No.1, pp. 116-122

[8] Bird, S. Natural Language Processing with Python – Analyzing Text with the Natural Language Toolkit // S. Bird, E. Klein, E. Loper / URL: http://www.nltk.org/book/

[9] Event-Based Programming. Taking Events to the Limit / URL: http://sharebookfree.com/event-based-programming-taking-events-to-the-limit/

[10] Myachykov, A. Visual Cues to Structural Choice in Visually Situated Sentence Production / A. Myachykov, D. Thompson, S. Garrod, C. Scheepers. – Frontiers in Psychology, Published online Jan 18, 2012 / URL: http://journal.frontiersin.org/Journal/10.3389/fpsyg.2011.00396/full9.

[11] Kallmeyer, L. Parsing Beyond Context-Free Grammars / Kallmeyer L. – Springer-Verlag Berlin Heidelberg 2010, – 247 p.

[12] Krämer, N.C. Nonverbal Communication / N.C. Krämer. – Human Behavior in Military Contexts, P. 150-188, / URL: http://books.nap.edu/catalog/12023.html

[13] Schuller, B., Batliner, A. Computational paralinguistics : emotion, affect and personality in speech and language processing / BjoËrn Schuller, Anton Batliner. – First Edition, John Wiley & Sons, Ltd , 2014, P. 321.

Computer paralinguistics texts contact languages

V.S. Krylov

Abstract. Information technology (IT) have formed a global infrastructure links between people. In the networks of the Internet emerged and are actively developing contacts languages. Syntax and grammar of these languages differs from the basic syntax and grammar of the language. It is like Pidgin and Creole. The Python programming language with a library of automated analysis of natural language NLTK provides an extensive set of tools for automatic analysis of language contact. Effective analyzer structure of texts of these languages - an event handler.

Keywords: *computer paralinguistics, event-oriented approach, krealizovanny language*