УДК 004.032.2

Н.С. Стёжка, Е.В. Шишкевич

Севастопольский национальный университет ядерной энергии и промышленности, Севастополь, Украина

ОЦЕНКА КАЧЕСТВА ДАННЫХ В СПЕЦИАЛИЗИРОВАННЫХ ИНФОРМАЦИОННО-ИЗМЕРИТЕЛЬНЫХ СИСТЕМАХ

Данная статья посвящена оценке возможности применения формализованного подхода к оценке качества технической документации, применяемой в информационно-измерительных системах. Наличие верно составленных документов является одной из составляющих успешной эксплуатации и оценки состояния технических объектов.

Ключевые слова: *удобочитаемость, качество технического текста, информационно-измерительная система.*

Введение

Качественная обработка измерительной информации в специализированных информационно-измерительных системах (например, атомных электростанциях и других сложных объектах) предусматривает использование технических текстов в виде цифробуквенных кодов-идентификаторов, названий контролируемых блоков, поясняющих записей, суждений о результатах контроля и диагностики, а так же в виде технической и эксплуатационной документации [1].

Любой технический текст, так или иначе связанный с данным видом систем должен быть однозначно понят во избежание принятия неверных управленческих решений и соблюдения норм безопасности.

Применение алгоритмов оценки удобочитаемости в технических текстах

Первую попытку оценки качества текстов с точки зрения удобочитаемости осуществил Б.А. Лайвли (B.A. Lively) и С.Л. Пресси (S.L. Pressey) из США [2]. Исследователи, столкнувшись с практической проблемой выбора учебников для студентов университетов, пришли к выводу, что книги были слишком перегружены научными и техническими терминами.

Это заставляло тратить большую часть времени на изучение терминологического словаря, нежели сущности изучаемых предметов.

По мнению учёных, основной причиной непонимания студентами учебного материала является выход технических терминов за пределы общеупотребительного словаря. А обращение к дополнительному словарю, дающему разъяснения, неизбежно вызывает увеличение объёма словаря и первоначального текста, причём его рост происходит по экспоненциальной зависимости [3].

Следующая попытка была сделана в 1952 г. профессором из штата Огайо, Р. Ганнингом (R. Gunning). В его работах была предложена формула следующего вида:

$$F_{Gunning} = 0,4 \cdot (w + b), \qquad (1)$$

где w – средняя длина предложения (соотношение общего количества слов к общему количеству предложений); b – процент длинных слов (при рекомендуемой длине отрывка текста в 100 слов – это фактически количество длинных слов в исследуемом отрывке текста), $b \le w$.

Наиболее сложные тексты, по его мнению, должны быть понятны студентам и выпускникам высших учебных заведений ($F_{Gunning} = 13 \div 17$), при этом для текстов общего употребления значение значения параметра удобочитаемости должно быть: $F_{Gunning} \le 8$. Не смотря на то, что тексты, исследованные Р. Ганнингом, не имели явной технической направленности, его методика легла в основу метрики Боэма, используемой для оценки качества текста комментариев к программному обеспечению [4].

Практически через 20 лет после опубликования работ Р. Ганнингом, вопросом удобочитаемости заинтересовалась армия США. Формула по оценке удобочитаемости FORCAST – это результат исследования которое проводилось с целью изучения требований к удобочитаемости текстов необходимых в ходе военных операций. С 1977 г. она применяется в рамках военно-воздушных сил США [5].

Данная формула имеет следующий вид:

$$F_{FORCAST} = 20 - \frac{N}{10}, \qquad (2)$$

где N – число однослоговых слов в образце текста длиной 150 символов.

Еще более популярной и широко применяемой

© Н.С. Стёжка, Е.В. Шишкевич

формулой стала формула разработанная Дж. Кинсайдом (J. Kincaid), офицером из штата Калифорния, в 1976 г. [6]. С его помощью формула удобочитаемости Р. Флеша (R. Flesch), была преобразована под нужды военно-морского флота США. Суть формулы сводится к оценке средней длины слов и предложений:

$$F_{\text{Flesch}-\text{Kincaid}} = 0,39 \cdot w + 11,8 \cdot p + 15,59, \qquad (3)$$

где p – средняя длина слова (общее количество слогов отнесенное к общему количеству слов).

Министерство обороны США применяет формулу Флеша-Кинсайда как стандартный тест на удобочитаемость для различных документов, с указанием, что документы, от которых зависит безопасность должны иметь значение показателя удобочитаемости не меньше 45 (чем ближе показатель к 100, тем более потенциально понятным является текст) [6].

В настоящее время широко используются различные методы оценки удобочитаемости текстовой информации, которые условно можно разделить на лексические и словарные методы. Наиболее распространенными являются лексические методы, определяющими показателями которых являются средняя длина слов (в слогах или в буквах) и предложений (в словах).

Эти параметры легко поддаются количественному определению с использованием современной компьютерной техники.

Однако, как показали результаты работ отечественных и зарубежных исследователей, многие алгоритмы оценки удобочитаемости могут быть применены к текстам для ограниченной группы лиц, к текстам определенной длины или же требуют соблюдения жестких условий при предварительной подготовке, поэтому в наши дни активно используется всего несколько формул удобочитаемости, например, формула Р. Флеша:

$$F_{\text{Flesch}} = 206,835 - 1,015 \cdot w - 84,6 \cdot p. \qquad (4)$$

В последние годы возможностью адаптации данного алгоритма заинтересовались отечественные учёные [7]. Адаптацию формулы Р.Флеша для украинского языка осуществил Партыко З.В. Его формула имеет следующий вид:

$$F_{\text{П}} = 206,835 - 5,952 \cdot w - 28,3 \cdot p. \qquad (5)$$

Адаптацию формулы Р.Флеша для русского языка осуществила Оборнева И.В. Ее формула имеет следующий вид:

$$F_{\text{O}} = 206,835 - 1,3 \cdot w - 60,1 \cdot p. \qquad (6)$$

Однако упомянутые авторы не используют информационный аспект удобочитаемости. В результате, при экспериментальной проверке параллельных текстов на близкородственных языках (русском и украинском) [8] позволили выявить расхождения в количественной оценке качества восприятия текста.

По нашему мнению это обусловлено отсутствием анализа информационного аспекта восприимчивости текста в зависимости от средней длины текста и количества непонятных, неверно прочитанных символов слова, невошедшего в обязательный словарь текста, выходящих за пределы общеупотребительного словаря.

Основной целью настоящей статьи является анализ возможности применения информационного подхода к оценке качества текстовой информации в специализированных информационно-измерительных системах.

Как следует из анализа выражений $(1) \div (6)$ все количественные оценки качества текста учитывают в основном среднюю длину предложения (w) и среднюю длину слова (p). Алгоритм Ганнинга дополнительно учитывает различие между длинами слов общеупотребительного языка и аномально длинными словами.

Многие исследователи отмечают, что непонятные (длинные) научно-технические термины образованы от слов латинского и древнегреческого происхождения, т.е. взяты из «мёртвых» языков. Эти слова, как правило, объединяются в группы научно-технических терминов длиной по 2-3 слова и часто заменяются аббревиатурами (например, арифметико-логическое устройство (АЛУ)).

С позиции информационной оценки длины высказывания все непонятные слова могут считаться равновероятными и давать максимальную энтропию (неопределенность).

В работах Р.Г. Пиотровского [3] введено понятие контекстной обусловленности текста, определение которой по смыслу и математическому выражению совпадает с понятием избыточности в классической теории информации:

$$R = 1 - \frac{H}{H_{\max}}, \qquad (7)$$

где $H_{\max} = \log p$, p – номер позиции символа в слове, т.е. длина слова.

В работе [9] предлагается оценивать уменьшение исходной энтропии слова с учётом зависимости рекомендованной Р.Г. Пиотровским:

$$H = H_{\max} \cdot e^{-S \cdot n}, \qquad (8)$$

где S – показатель уменьшения энтропии; n – номер позиции символа в сообщении (слове, предложении), т.е. фактически длина слова, предложения.

Воспользовавшись выражением (7) формулу

(8) можно представить в виде:

$$R = 1 - e^{-S \cdot n}. \qquad (9)$$

Величина (1 – R) имеет смысл коэффициента сжатия информации:

$\mu = e^{-S \cdot n}$, при $n \cdot S = 1$ имеем $S = \frac{1}{n}$, $\mu = \frac{1}{e}$.

Округлив значение до 0,4 можно получить эмпирический коэффициент возможного уменьшения длины цифробуквенного кода (текстового сообщения), принятый в работах Ганнинга и других авторов.

Вполне возможно, что в его работах $e^{-1} = 0{,}367$ была округлена до эмпирического коэффициента 0,4, но в проанализированных авторами публикациях нет никаких сведений по данному вопросу. Тем более, тексты, исследованные Р. Ганнингом, были англоязычными и не имели явной технической направленности.

При значении показателя $S = \frac{1}{n}$, где n – экспериментально установленное или заданное среднее количество лексических элементов (символов в слове, слов в предложении) в тексте сообщения на экране дисплея информационно-измерительной системы (ИИС) рост избыточности технического текста до расчётного уровня 0,63 происходит по разным экспонентам.

Результаты проведенных в соответствии с [9] лабораторных работ со студентами третьего курса позволяют уверенно различить отрывки технического и художественного текстов на русском, украинском и английском языках.

Выводы

Таким образом можно полагать, что оценку удобочитаемости текстовых сообщений по методу Ганнинга, с учётом предложенных рекомендаций, можно использовать для оценки документации применяемой в современных ИИС.

Список литературы

1. Цапенко М.П. Измерительные информационные системы: Структуры и алгоритмы, системотехническое проектирование / М.П. Цапенко. – М.: Энергоатомиздат, 1985. – 438 с.

2. William H. DuBay The Classic Readability Studies / H. William. – Impact Information Costa Mesa, 2006. – 112 p.

3. Пиотровский Р.Г.. Математическая лингвистика / Р.Г. Пиотровский, К.Б. Бектаев, А.А. Пиотровская. – М.: Высшая школа, 1977. – 383 с.

4. Боэм Б. Инженерное проектирование программного обеспечения: пер. с англ. / Б. Боэм. – М.: Мир, 1985. – 512 с.

5. The technique of clear writing by Gunning, R. McGraw-Hill International Book Co; New York, NY, 1952.

6. Рогушина Ю.В. Использование критериев удобочитаемости текста для поиска информации, соответствующей реальной потребности пользователя / Ю.В. Рогушина // Проблеми програмування. – 2007. – № 3. – С. 76-87.

7. Шишкевич Е.В. Комплексный вероятностно-информационный показатель оценки качества научно-технических текстов на естественных языках / Е.В. Шишкевич, Н.С. Стёжка // Сборник научных трудов СНУЯЭиП. – Севастополь: СНУЯЭиП, 2010. – № 3(35). – С. 162-167.

8. Попова Я.И. Стандартизация учебной литературы средней школы по критерию удобочитаемости / Я.И. Попова, Е.В. Шишкевич // Научные ведомости БелГУ. – 2010. – № 12 (83). – С. 142-147.

9. Методические указаний к выполнению лабораторных и практических работ по дисциплине "Метрология, стандартизация, сертификация и аккредитация" / Сост. Е.В. Шишкевич, М.А. Лебедева – Севастополь: СевНТУ, 2004. – 27 с.

Поступила в редколлегию 20.08.2011

Рецензент: д-р техн. наук, проф. Ю.П. Мачехин, Харьковский национальный университет радиоэлектроники, Харьков, Украина.

ОЦІНКА ЯКОСТІ ДАНИХ
В СПЕЦІАЛІЗОВАНИХ ІНФОРМАЦІЙНО-ВИМІРЮВАЛЬНИХ СИСТЕМАХ

Н.С. Стьожка, Е.В. Шишкевич

Дана стаття присвячена оцінці можливості вживання формалізованого підходу до оцінки якості технічної документації, яка використовується в інформаційно-вимірювальних системах. Наявність вірно складених документів є однією із складових успішної експлуатації і оцінки стану технічних об'єктів.

Ключові слова: *легкість для читання, якість технічного тексту, інформаційно-вимірювальна система.*

DATA QUALITY ESTIMATION
IN SPECIALIZED INFORMATION-MEASURING SYSTEMS

N.S. Styozhka, E.V. Shishkevich

Given article is devoted to the estimation of possibility to apply the formalized approach to the quality estimation of the technscal documentation which is applied in information-measuring systems. Presence of truly made documents is one of the components of successful operatiand an estimation of the technical objects.

Keywords: *readability, quality of the technical text, information-measuring system.*