«Сейфуллин оқулары – 12: Ғылым жолындағы жастар-болашақтың инновациялық әлеуеті» атты Республикалық ғылыми-теориялық конференция материалдары = Материалы Республиканской научно-теоретической конференции **«Сейфуллинские чтения-12:** Молодежь в науке - инновационный потенциал будущего" . – **2016**. – Т.1, ч.3 – С.287-289

АНАЛИЗ ЭНТРОПИЙНОЙ ИЗБЫТОЧНОСТИТЕХНИЧЕСКОГО ТЕКСТА НА КАЗАХСКОМ ЯЗЫКЕ

Мирманов А.Б., Сарсембиева Э.К.

Увеличение казахскоязычных Интернет-ресурсов, перевод на казахский язык и создание единого словаря веб-терминов повышает значимость подготовки специалистов информационных технологий со знанием технического казахского языка. Внедрение в ВУЗах Казахстана дисциплины «Профессиональный казахский язык» способствует расширению сферы применения государственного языка. Подготовка выпускников специальностей, связанных с инфокоммуникационными технологиями, имеет свою специфику в том числе и в обучении технической терминологии и технического языка. Развитие казахского языка требует от будущего инженера знание делопроизводства на государственном языке.

Большинство слов, используемых в сфере инфокоммуникаций, имеют англоязычные корни или заимствованы из других языков, что делает технический текст на казахском языке понятным для восприятия квалифицированными специалистами. Но, в последнее время, отечественные языковеды переводят на казахский язык большое количество международных терминов, тем самым появляются новые слова и словосочетания.

С развитием электронного документа оборота, веб-ресурсов государственных и частных компаний, управление и решения о дальнейших шагах принимается на основе получаемой из них информации. Информационныепотоки пересылаемые по каналам связи постоянно растут. Большая их часть приходится на текстовые данные. Причем в условиях двуязычья тесты идут на казахском и русском языках, замедляя процесс, передачи информации от источника к получателю. При этом этот факт связан не только с необходимостью формировать два сообщения, то есть производить дословный ил смысловой перевод, но и само сообщении становиться большим как минимум в два раза. Использование двух алфавитов при формировании сообщений, увеличивает количество информации генерируемой источником дискретных сообщений и снижает скорость передачи данных.

В предлагаемом исследовании информационных потоков анализ проводился на примере двух технических текстов, объемом около 25 страниц для каждого языка.

При проведении анализа было принято, что для русского языка алфавит сообщения содержит 32 символа: 33 буквы и пробел, при этом буквы е,ё и буквы ь,ъ приняты за один соответственно. Для казахского кириллического алфавита - 41 символ: 42 буквы и пробел, из которых 33 буквы русского алфавита (буквы е,ё и буквы ь,ъ приняты за один соответственно) и 9 специфических букв казахского языка.

Оценка избыточности и энтропийной эффективности позволяет изучать особенности различия языка и вести работу по уменьшению информационных помех.

Для определения эффективности и избыточности текстовой информации, содержащей терминологию по специальности «Радиотехника, электроника и телекоммуникаций» составленной на казахском и русском языках были определены количество информации и суммарная неопределенность информационного текста. Проведен расчет коэффициента статического сжатия и относительной эффективности.

Энтропия $H(X)$ и количество информации I зависят от исходного количества рассматриваемых вариантов N и априорных вероятностей реализации каждого из них $P(x)$.При этом максимальная энтропия определяется по формуле [1], а при использовании равномерного кодирования - средняя длина кода равна энтропии источника.

$$H_{max} = \log_2 N \qquad [1]$$

Для казахского и русского языка, соответственно максимальная энтропия будет 5,36 бит и 5 бит, при равномерном кодировании 6 бит и 5 бит, для кода ASCII по 8 бит.

Считая, что появление букв в слове является независимым событием, расчет энтропии проведем по формуле Шеннона:

$$H(x) = -\sum_{i=0}^{n} p(x_i) \times \log_2 p(x_i) \qquad [2]$$

Для нахождения энтропии технического текста найдем частоты и вероятность появления букв. Расчет реальной энтропии показал: $H_{каз}(X)$=4,58 бит; $H_{рус}(X)$=4,46 бит.

Следовательно, среднее количество информации, переносимой одним двоичным элементом комбинации, согласно формуле: $H(x)/H_{max}$, будет равно для казахского языка – 0,854 (реальная), 0,654 (при равномерном КОИ-7), 0,572 (ASCII) и для русского – 0,891 (реальная), 0,891 (при равномерном МТК-2), 0,556 (ASCII).

Получается, что казахский язык эффективен на 85,5% русский – на 89,1%.

Таким образом, средняя недогрузка (избыточность) каждого двоичного элемента$D = 1 - H(x)/H_{max}$составит:

$D_{каз}=0,145;$ для ASCII$D_{каз}=0,428$

$D_{рус}$*=0,109;* для ASCII$D_{каз}$*=0,443*

Интересным моментом является избыточность для языков при разных способах кодирования.

До сих пор анализ проводился при условии независимых событий, т.е. не учитывалась специфика языка и условные вероятности букв в техническом тексте. Условную энтропию первого порядка для алфавита, где известны вероятности появления одной буквы после другой, найдем по формуле [2]:

$$H(S) = -\sum_i p_i \sum_j p_i(j) \times \log_2 p_i(j)$$[3]

В анализе для вероятностных событий подряд идущих взаимосвязанных букв, очевидно, эффективность казахского технического текста будет убывать, если вычислять ее по более длинным последовательностям.

В докладе будет раскрыт представленный анализ, показано сравнение энтропий технического текста английскогои казахского языков.Также проведен анализ международной терминологии и новых переведенных на казахский язык заимствованных слов.

Список литературы

1. Шеннон К. Математическая теория связи. (Shannon C.E. A Mathematical Theory of Communication. BellSystemTechnicalJournal. -1948. - Т. 27. - С. 379-423, 623–656.)

2. Шеннон К. Работы по теории информации и кибернетике. –М.: Изд. иностр. лит., 2002.