КІБЕРНЕТИКА

Ю.Г. КРИВОНОС, Ю.В. КРАК, А.В. БАРМАК, Р.А. БАГРИЙ

УДК 004.93.12

ПРЕДИКТИВНАЯ СИСТЕМА НАБОРА ТЕКСТА ДЛЯ УКРАИНСКОГО ЯЗЫКА

Аннотация. Исследована интеллектуализация ввода информации с помощью системы ускоренного ввода текста в цифровые устройства в целях построения модели корпуса разговорного украинского языка и системы набора текста, базирующейся на этой модели. Такая система использует меньшее количество команд для ввода букв и прогнозирует варианты слов, основываясь на данных корпуса слов и словосочетаний для общения. Экспериментально показано, что для построенного корпуса достаточно эффективен ввод текста с помощью четырех и шести клавиш-команд.

Ключевые слова: альтернативная коммуникация, формирование корпуса слов, N-граммы, прогнозирование.

ВВЕДЕНИЕ И ПОСТАНОВКА ЗАДАЧИ

Вовлечение в общественную жизнь людей с постоянными или временными проблемами, связанными с утраченной способностью речевого общения, обусловило разработку концепции дополнительной и альтернативной коммуникации (Augmentative and Alternative Communication, AAC) [1].

Такая AAC использует спектр разнообразных способов, которые помогают людям высказывать свои мысли и эффективно общаться. Современные системы AAC представлены, в том числе, и многофункциональными средствами коммуникации на основе сложных технических устройств: сенсорных экранов, синтезаторов языка и др. Они обеспечивают значительное расширение словаря, позволяют задавать тему беседы, объединять одновременно несколько тем, общаться на расстоянии, а также упрощают коллективное общение и по телефону. Современное развитие вычислительной техники и информационных технологий способствует существенному усовершенствованию этих средств AAC.

В настоящей работе предложена информационная технология AAC [2, 3], аппаратно-программная реализация которой должна обеспечить коммуникацию максимально возможными способами, при этом важное значение имеет интеллектуализация ввода информации.

Постановка задачи: исследовать интеллектуализацию ввода информации с помощью системы ускоренного ввода текста в цифровые устройства в целях построения модели корпуса разговорного украинского языка (устной речи) и основанной на этой модели предиктивной системы набора текста. Последняя использует меньшее количество команд для ввода букв и прогнозирует варианты слов, базируясь на данных корпуса слов и словосочетаний для общения.

МОДЕЛИРОВАНИЕ СИСТЕМЫ ВВОДА ТЕКСТА С ПРОГНОЗИРОВАНИЕМ

Алгоритмы прогнозирования способны автоматически завершать ввод текста, что позволяет оптимизировать время его введения. Для прогнозирования необходимо найти баланс между скоростью ввода и функциональностью. Чем больший словарь и чем больше используется технологий ввода, тем меньше становится время отклика, но повышается качество результатов.

Для исследования и моделирования системы ввода текста с прогнозированием предлагается следующий подход, включающий:

— формирование множества (корпуса) слов (словосочетаний) украинского языка (ограниченного словами повседневного общения) с построением соответствующей модели прогнозирования употребления слов в словосочетании;

— группирование (с соответствующим кодированием) множества букв украинского алфавита в определенном порядке следования (алфавитный, клавиатурный (qwerty), по частоте встречаемости и т.д.).

Для формирования корпуса слов предлагается применять экспертный подход. Это обусловлено тем, что нужно подобрать слова и словосочетания повседневного общения. Для этого используются источники из контента украиноязычных сайтов, периодической печати, словарей-разговорников и т.д.

Существует несколько известных корпусов украинского языка:

— Национальный корпус украинского языка (НКУМ) [4], содержащий тексты письменного и устного (разговорного) стиля национального языка;

— Украинский национальный лингвистический корпус (УНЛК) [5], созданный Украинским языково-информационным фондом (УМИФ) и насчитывающий более 43 млн. слов;

— корпус украинского языка на лингвистическом портале mova.info [6], предоставляющий пользователям возможность поиска слов по нескольким подкорпусам, а именно художественной прозе, научным, поэтическим, фольклорным, законодательным и публицистическим текстам.

В большинстве имеющихся языковых корпусов сохраняются только письменные тексты различных стилей, а формирование подкорпуса разговорного стиля лишь декларируется. Это связано с тем, что собирание текстов разговорного стиля требует большого количества интервьюеров и респондентов для записи аудиоматериала, а расшифровка собранных записей достаточно трудоемка.

Существует и иной вариант — это создание корпуса текстов языка общения в сети Интернет, который используют при коммуникации (переписке) в социальных сетях, чатах, по электронной почте, на форумах и т.д. Этот способ собирания текстов довольно быстрый, но имеет недостатки: во-первых, во время Интернет-коммуникации собеседники не видят друг друга, во-вторых, при написании часто используются сокращения, которые в реальной беседе не встречаются, в-третьих, различие в уровне знаний участников общения приводит к большому количеству ошибок в текстах и использованию заимствованных слов из других языков. Для устранения этих недостатков требуется значительная ручная проверка текстов, что трудоемко для больших объемов данных.Таким образом, использовать существующие корпусы для реализации коммуникации между людьми невозможно и возникает необходимость в формировании ограниченного подкорпуса разговорного украинского языка.

Для решения этой задачи решено использовать диалоги на бытовые темы из словарей-разговорников иностранных языков. Такие диалоги моделируют разговор между людьми, которые видят друг друга, а также в наиболее возможных бытовых ситуациях с использованием ограниченного набора слов и фраз. Для системы коммуникации людей с ограниченными возможностями подобные свойства диалога наиболее приемлемы.

Как отмечалось ранее, для сбора текстов используются разговорники, учебники, пособия и другие Интернет-ресурсы, содержащие диалоги. Для последующего формирования модели полученные диалоги следует разбить на базовый и тестовые блоки в целях определения достаточной наполненности корпуса для задачи прогнозирования слов и словосочетаний при вводе начальных букв.

В алгоритмах обработки естественного языка (Natural Language Processing, NLP) [7] принято использовать следующие понятия: типы (types) — различные слова и токены (tokens) — все слова. В обоих случаеях это последовательности слов в тексте.

Для построения модели на базе полученных текстовых наборов их необходимо обработать. В настоящее время существует множество способов разбить электронный текст на отдельно значимые единицы для их последующей компьютерной обработки. Для решения поставленной задачи предлагается каждый набор токенизировать [7], т.е. разбить текст вначале на предложения, а потом на осмысленные элементы (слова, фразы, символы), называемые токенами. Во время токенизации из текста убираются все символы, которые не относятся к украинскому алфавиту. Алгоритм токенизации текста предложено реализовать с помощью набора регулярных выражений [8].

Одним из способов стандартизации есть регулярные выражения — язык для задания текстового поиска строк. Формально регулярное выражение является алгебраическим обозначением, которое характеризует набор строк. Регулярные выражения облегчают обработку значительных текстовых объемов с помощью небольших шаблонов для поиска. Регулярная функция выражения поиска, применяемая к корпусу, возвращает все тексты, соответствующие шаблону.

Для моделирования полученного объема текстовой информации в целях последующего прогнозирования слов в словосочетаниях предлагается полученные предложения из диалогов разбить на N-граммы: юниграммы, биграммы и триграммы [7] (рис. 1).

Для рассматриваемой задачи целью построения N-грамм (последовательностей из N слов) моделей является определение вероятности использования заданного слова или фразы (словосочетания), например, биграммы: «я хочу», «мені потрібно», триграммы: «у мене болить», «я хочу знати».

При обработке естественного языка N-граммы предназначены в основном для предсказания на базе вероятностных моделей. В N-граммной модели рассчитывается вероятность использования последнего слова N-граммы, если известны все предыдущие. При применении этого подхода для моделирования языка предполагается, что появление каждого слова зависит только от предыдущих слов.

Целью построения N-граммных моделей есть определение вероятности использования заданной фразы (словосочетания). Эту вероятность можно задать

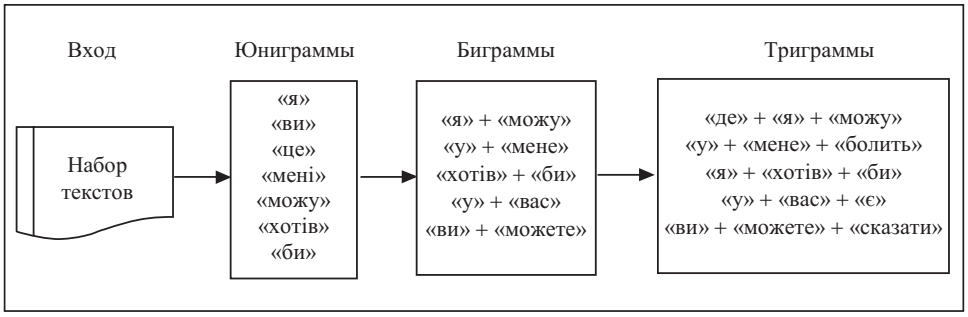

Рис. 1

формально как вероятность возникновения последовательности слов в некотором корпусе (наборе текстов). Для оценки этих вероятностей необходим соответствующий метод. Предлагается для этого использовать один из самых простых и наиболее интуитивно понятных способов оценки достоверности — метод максимального правдоподобия (maximum likelihood estimation, MLE) [7].

Для оценки максимального подобия слов нужно принять модель — параметры, максимизирующие достоверность этого сходства для заданных слов. Для MLE оценку параметров модели N-граммы можно получить как нормализованное количество от корпуса. Последний представляет собой набор текстов, который статистически репрезентативен для моделирования языка. Например, можно оценить биграмм-вероятность слова w_n, учитывая предыдущее слово w_{n-1}, подсчитывая вхождение биграмм $C(w_n, w_{n-1})$ и нормируя суммой всех биграмм, содержащих первое слово w_n:

$$P(w_n | w_{n-1}) = \frac{C(w_{n-1}, w_n)}{\sum C(w_{n-1}, w)}. \tag{1}$$

Так как количество всех биграмм, начинающихся со слова w_{n-1}, равно количеству юниграмм для этого слова w_{n-1}, выражение (1) упрощается:

$$P(w_n | w_{n-1}) = \frac{C(w_{n-1}, w_n)}{C(w_{n-1})}. \tag{2}$$

В общем случае N-граммной модели формула для оценки параметров MLE имеет вид

$$P(w_n | w_{n-N+1}^{n-1}) = \frac{C(w_{n-N+1}^{n-1}, w_n)}{C(w_{n-N+1}^{n-1})}. \tag{3}$$

Модели N-грамм в основном предназначены для предсказания слова, так как они эффективны и просты в использовании. К тому же статистику частот, которая применяется для расчета MLE оценки, можно получить непосредственно из текстов без особых усилий.

В настоящей работе предложена адаптированная технология для ввода информации меньшим количеством клавиш [3]. Она предполагает для ввода использовать блоки, состоящие из сгруппированных букв украинского алфавита в определенном порядке следования: алфавитный, клавиатурный (qwerty), по частоте встречаемости и т.д. Согласно данной технологии одним нажатием клавиши будет выбрана не конкретная буква, а их набор, зависящий от текущей раскладки символов по блокам. Так, для того чтобы набрать слово «болить», необходимо выбрать блоки символов в порядке, показанном на рис. 2. Введенное таким образом слово представляет собой условный код, полученный для текущей раскладки символов на клавиатуре. Далее этот код сравнивается с внутренним словарем, который содержит слова или словосочетания (N-граммы), и предлагается наиболее вероятное слово.

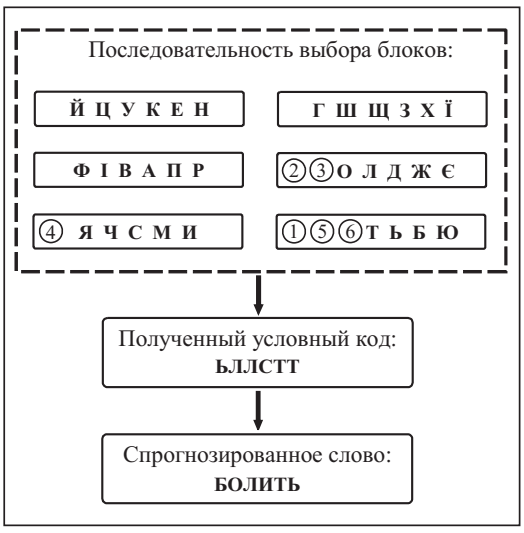

Рис. 2

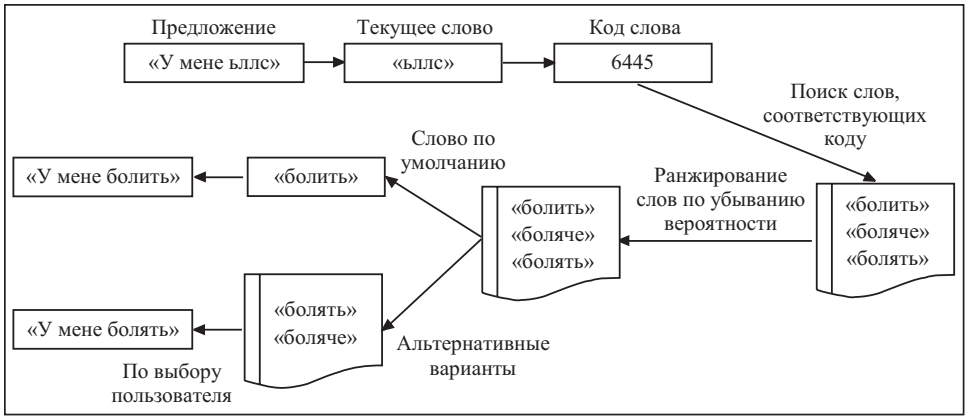

Рис. 3

Классический подход к прогнозированию по введенным буквам в этом случае усложняется тем, что при использовании предложенной технологии быстрого ввода на вход подаются слова, зашифрованные кодом, соответствующим выбранному порядку следования букв в блоках клавиатуры. В этом случае одному коду может соответствовать несколько слов. Для решения данной проблемы предлагается особый порядок для прогнозирования. Пример работы такого подхода приведен на рис. 3.

Суть предложенного подхода заключается в использовании N-граммной модели для прогнозирования наиболее вероятного вводимого слова.

Для комфортной работы с системой необходимо, чтобы как можно больше вводимых слов прогнозировались в качестве слов по умолчанию, тогда не нужны дополнительные действия пользователя. Если необходимое слово находится ниже в списке слов для ввода, то его выбор требует одного действия в каждой позиции в списке. Отметим, что можно спрогнозировать только те слова, которые имеются в словаре. Если слово в словаре не значится, то для его ввода необходимо затратить гораздо больше времени, так как его нужно ввести полностью по буквам.

Прогнозирование слов реализовано следующими моделями:

— триграммной, когда введенное предложение состоит из трех или более слов. Для расчета триграмм-вероятности используются два предыдущих слова, например, для фразы «у мене болить» вероятность будет записана как $P(\text{болить} \mid \text{у мене})$;

— биграммной для словосочетания из двух слов или в случае, когда выражение не найдено среди триграмм. При этом для расчета вероятности выбирается только предыдущее слово: $P(\text{болить} \mid \text{мене})$;

— юниграммной, когда прогнозируемое слово первое в предложении (т.е. не существует предыдущего слова) или словосочетание не найдено среди биграмм или триграмм. Вероятность слова будет равна его нормализованной частоте $P(\text{болить})$.

РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТАЛЬНОГО ТЕСТИРОВАНИЯ

Для моделирования корпуса диалогов на украинском языке собран набор, состоящий из более 400 диалогов на различные темы, общий объем которых приблизительно 20000 фраз и 100000 слов.

Этот набор текстов разбит на 13 блоков: один базовый (начальный) и 12 тестовых для контроля за узнаваемостью текста при прогнозировании. Общая информация о каждом блоке представлена в табл. 1.

Таблица 1

Название блока	Количество		
	фраз	слов	различных слов
Базовый	2395	9028	2661
Тестовый-1	1534	7950	2675
Тестовый-2	1825	7387	2061
Тестовый-3	1290	8447	2690
Тестовый-4	1451	8314	2643
Тестовый-5	1305	8224	2674
Тестовый-6	1195	8228	2818
Тестовый-7	1195	7934	2701
Тестовый-8	1179	8135	2614
Тестовый-9	1234	8248	2705
Тестовый-10	1327	8342	2708
Тестовый-11	1612	7315	2163
Тестовый-12	1281	7166	2229

В результате в базовой модели N-грамм содержится 2661 типов юниграмм (различных слов), 6527 биграмм и 8071 триграмм. Базовая N-граммная модель постепенно расширялась добавлением новых текстов из тестовых блоков. Модель, построенная для всего набора текстов, содержит 15065 типов юниграмм, 65386 биграмм, 93371 триграмм. Так как таблица юниграмм представляет перечень различных слов с их частотой встречаемости в текстах, то можно считать, что размер словаря составляет 15000 слов.

На каждом шаге анализировались тестовые блоки, которых еще не было в текущей N-граммной модели. Это позволило проследить, как происходило построение (наполнение) данной модели.

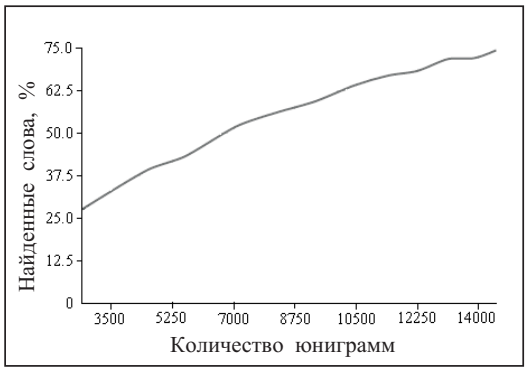

Рис. 4

На рис. 4 показан график узнаваемости текста, т.е. процентное отношение количества различных известных слов (типов) к каждому этапу расширения N-граммной модели. Так, для базовой модели в тестовых блоках известна только четвертая часть слов, для 7000 юниграмм количество известных и неизвестных слов в текстах одинаково. После формирования всей N-граммной модели три четверти различных слов произвольного текста стали известными.

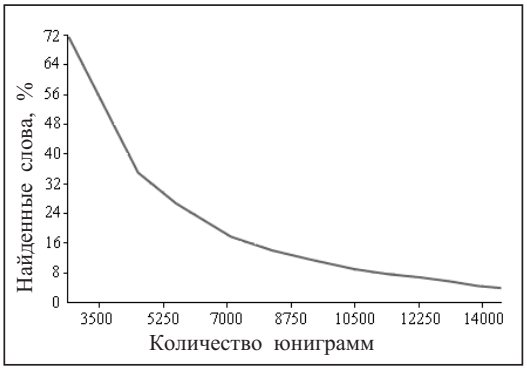

Рис. 5

На рис. 5 показан график прироста новых слов (типов) в N-граммной модели для каждого добавленного нового тестового блока. Видно, что в последних блоках увеличивался размер модели менее чем на 4 %.

Таблица 2

Название блока	Количество низкочастотных юниграмм	Узнаваемость	
		текста, %	известных слов, %
Базовый	738	91.83	72.27
Тестовый-1	798	89.96	70.17
Тестовый-2	400	94.59	80.59
Тестовый-3	632	92.52	76.51
Тестовый-4	607	92.70	77.03
Тестовый-5	640	92.22	76.07
Тестовый-6	735	91.07	73.92
Тестовый-7	677	91.47	74.94
Тестовый-8	633	92.22	75.78
Тестовый-9	710	91.39	73.75
Тестовый-10	618	92.59	77.18
Тестовый-11	503	93.12	76.75
Тестовый-12	498	93.05	77.66

Анализ юниграмм (см. рис. 4, 5) показал, что примерно половина всех слов (токенов) корпуса имеет высокую частоту встречаемости (более 100 раз на 100000 слов). Примерно 55 % типов юниграмм имеют частоту, равную единице, т.е. они встречаются один раз в данном наборе текстов. В табл. 2 показано, сколько таких низкочастотных юниграмм имеется в каждом тестовом блоке.

Как видно из табл. 2, общая узнаваемость текста без учета низкочастотных юниграмм составляет в среднем 92 %, а количество известных токенов в тестовых блоках в среднем равно 75 %.

Итак, можно предположить, что в каждом новом тестовом блоке будет содержаться от 20 до 30 % новых различных слов, которые не встречались ранее. Большая часть из них будет иметь низкую частоту встречаемости, что не позволит улучшить качество предсказания. Поэтому дальнейшее накопление текстов и расширение N-граммной модели неэффективно. Лучшим вариантом является создание отдельной персональной N-граммной модели, где будут содержаться те слова и последовательности слов, которые чаще употребляются конкретным человеком.

Для определения качества прогнозирования проведен ряд экспериментов для каждой модели. Тестировались шесть блоков клавиатурного порядка следования букв и четыре блока частотной последовательности букв с учетом гласных и согласных [3]. При этом количество типов юниграмм составляло приблизительно 15000, биграмм — 65000, триграмм — 93000.

Первоначальное тестирование заключалось в прогнозировании того же текста, на основе которого были сформированы N-граммные модели. Общее количество фраз для прогнозирования — более 18000.

Юниграммная модель показывала низкое качество прогнозирования: 87.5% для шести блоков клавиш (рис. 6, *а*) и 83.5 % для четырех блоков (рис. 6, *б*) распределения. Для других случаев необходимо применять одно дополнительное действие для выбора слова, а иногда два и более таких действий. Для биграммной модели этот показатель значительно больше и практически одинаковый для обоих случаев: 95.5–96%. Этого вполне достаточно для комфортной работы. Триграммная модель еще немного улучшает прогнозирование, и качество почти не зависит от количества блоков.

Последующие тестирования заключались в прогнозировании произвольного текста. Объем текста составил приблизительно 3000 фраз и 15000 слов. Для

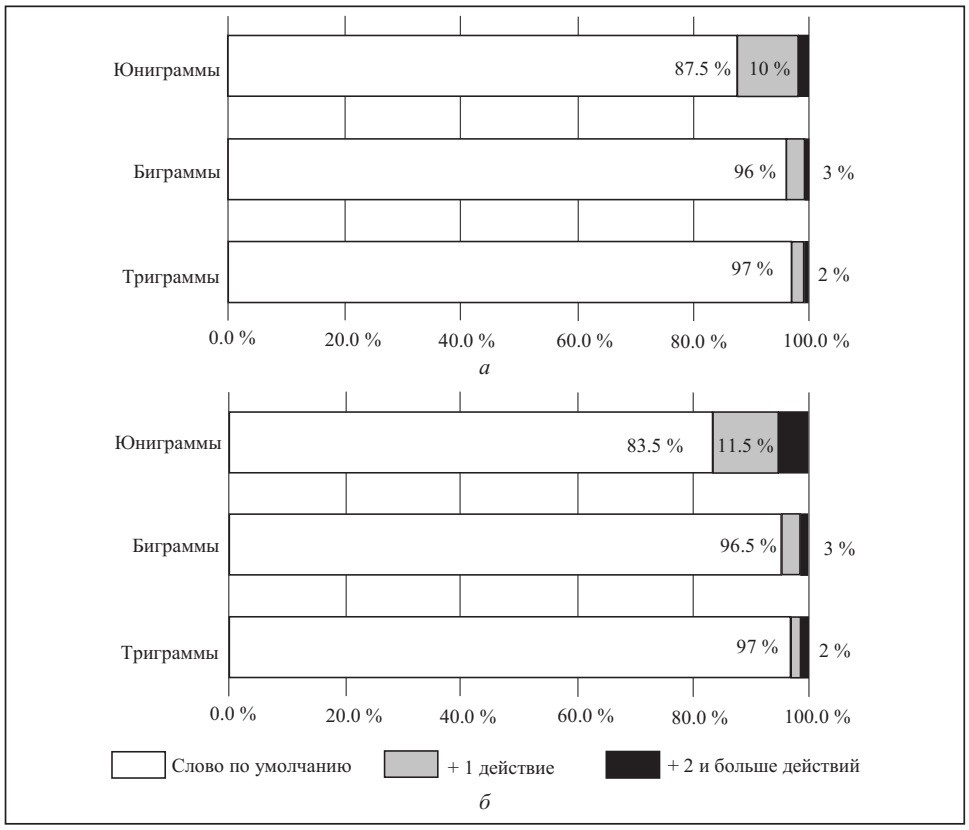

Рис. 6

шести блоков правильно спрогнозировано 90 % слов, известных N-граммной модели. Для четырех блоков качество прогнозирования составило 89 %. Снижение точности прогнозирования можно объяснить тем, что в произвольном тексте всегда имеется определенная часть новых типов биграмм и триграмм. Также в случае отсутствия слова в N-граммной модели, следующее за ним слово можно спрогнозировать только с помощью юниграммной модели.

ЗАКЛЮЧЕНИЕ

В статье для реализации информационной технологии альтернативных подходов к общению [2] предложена предиктивная система набора текста. Для украинского языка создан корпус разговорного контента на основе N-граммной модели. Описаны подходы для быстрого ввода разговорных диалогов с помощью ограниченного набора клавиш.

Дальнейшие исследования направлены на реализацию предложенного способа альтернативного общения с помощью стандартных гаджетов (планшеты, телефоны) для организации диалогов с людьми, у которых временно отсутствует или затруднен канал основной вербальной коммуникации. Поскольку приведенный подход является общим, будут изучаться возможности его использования для других языков.

СПИСОК ЛИТЕРАТУРЫ

1. Augmentative and Alternative Communication (AAC). URL: http://www.asha.org/public/speech/disorders/AAC/.
2. Кривонос Ю.Г., Крак Ю.В., Бармак А.В., Багрий Р.А. Новые средства альтернативной коммуникации для людей с ограниченными возможностями. *Кибернетика и системный анализ.* 2016. Т. 52, № 5. С. 3–13.

3. Крак Ю.В., Бармак А.В., Багрий Р.А., Стеля И.О. Система ввода текста для альтернативной коммуникации. *Проблемы управления и информатики*. 2017. № 1 С. 128–137.
4. Демська-Кульчицька О. Національний корпус української мови: концептуальний аспект. *Лексикографічний бюлетень.* Київ: Ін-т української мови НАН України, 2006. Вип. 13. С. 5–9.
5. Сидорчук Н.М. Организация данных и функциональная структура лексикографической системы «Украинский национальный лингвистический корпус». *Математичні машини і системи*. 2006. № 2. С. 126–135.
6. Дарчук Н. Дослідницький корпус української мови: основні засади і перспективи. *Вісник Київського національного університету імені Тараса Шевченка*. 2010. Вип. 21. С. 45–49.
7. Jurafsky D., Martin J.H. Speech and language processing. 2nd ed. New Jersey: Prentice Hall, 2008. 1024 p.
8. Bird S., Klein E., Loper E. Natural language processing with Python. Sebastopol, CA (USA): O'Reilly Media, 2009. 504 p.

Надійшла до редакції 22.02.2017

Ю.Г. Кривонос, Ю.В. Крак, О.В. Бармак, Р.О. Багрій

ІНТЕЛЕКТУАЛЬНА СИСТЕМА НАБОРУ ТЕКСТУ ДЛЯ УКРАЇНСЬКОЇ МОВИ

Анотація. Досліджено інтелектуалізацію введення інформації за допомогою системи прискореного введення тексту в цифрові пристрої з метою побудови моделі корпусу розмовної української мови та системи набору тексту, яка базується на цій моделі. Така система використовує меншу кількість команд для введення букв та прогнозує варіанти слів, базуючись на даних корпусу слів та словосполучень для спілкування. Експериментально показано, що для побудованого корпусу достатньо ефективним є введення тексту за допомогою чотирьох та шести клавіш-команд.

Ключові слова: альтернативна комунікація, формування корпусу слів, N-грами, прогнозування.

Iu.G. Kryvonos, Iu.V. Krak, O.V. Barmak, R.O. Bagriy

PREDICTIVE TEXT TYPING SYSTEM FOR THE UKRAINIAN LANGUAGE

Abstract. The paper investigated intellectualization of text input by a predictive text input system in order to construct a model of the corps of spoken Ukrainian language and text typing system based on this model. This system uses fewer instructions to input letters and predicts versions of words based on the corpus of words and phrases for communication. It is shown experimentally that text input using 4 and 6 command keys is very efficient for the constructed corps.

Keywords: alternative communication, formation of the corpus of words, N-grams, prediction.

Кривонос Юрий Георгиевич,
академик НАН Украины, доктор физ.-мат. наук, профессор, заместитель директора Института кибернетики им. В.М. Глушкова НАН Украины, Киев.

Крак Юрий Васильевич,
доктор физ.-мат. наук, профессор, заведующий кафедрой Киевского национального университета имени Тараса Шевченко, старший научный сотрудник Института кибернетики им. В.М. Глушкова НАН Украины, Киев, e-mail: yuri.krak@gmail.com.

Бармак Александр Владимирович,
доктор техн. наук, профессор кафедры Хмельницкого национального университета,
e-mail: alexander.barmak@gmail.com.

Багрий Руслан Александрович,
старший преподаватель кафедры Хмельницкого национального университета,
e-mail: gcardinal2009@gmail.com.