УДК 001
ББК 1
Н34

Р е ц е н з е н т ы: д-р ист. наук, проф. А.Ю.Дворниченко,
д-р. физ.-мат. наук, проф. В.Ю.Тертычный-Даури
О т в. р е д а к т о р: канд. физ.-мат.наук, доц. В.Г.Быков

Наука СПбГУ — 2021.
Сборник материалов Всероссийской конференции по естественным и гуманитарным наукам с международным участием, 28 декабря 2021 года. СПб: Свое издательство, 2022. — 1132 с.

ISBN 978-5-4386-2169-0

Сборник содержит материалы докладов Всероссийской конференции по естественным и гуманитарным наукам с международным участием **«Наука — 2021»**, проходившей 28 декабря 2021 г. в Санкт-Петербургском государственном университете в цифровом формате. В сборнике представлены результаты теоретических и прикладных исследований по самому широкому кругу актуальных проблем в областях естественных и точных наук (биология, математика, механика, информатика, медицина, науки о Земле, физика и астрономия, химия), а также социальных и гуманитарных наук (искусство, история, международные отношения и политология, науки о языках и литература, психология, педагогика, когнитивные науки, социология, журналистика и массовые коммуникации, философия, конфликтоведение, этика, культурология, религиоведение, экономика и менеджмент, юридические науки).

Междисциплинарный характер материалов сборника позволяет адресовать его ученым всех областей знания, а также использовать в научной, учебной и учебно-методической работе преподавателей высших учебных заведений.

Материалы докладов в сборнике представлены в авторской редакции.

Венцов А. В.[1], Риехакайнен Е.И.[1]

ОБ ОПЫТЕ РАЗРАБОТКИ АЛГОРИТМА РАСПОЗНАВАНИЯ РЕДУЦИРОВАННЫХ СЛОВОФОРМ НА МАТЕРИАЛЕ РУССКОЙ УСТНОЙ РЕЧИ

Несмотря на значительные достижения в области автоматического распознавания звучащей речи, ни одна из существующих автоматических систем до сих пор не справляется с задачей распознавания непринужденной естественной звучащей речи так же эффективно, как носитель языка. Одной из главных сложностей при этом остается фонетическая редукция словоформ. Цель нашего исследования — разработать алгоритм взаимодействия различных языковых уровней в процессе восприятия речи, приближенный к тому, каким образом распознает естественный речевой сигнал человек. Постановка такой задачи стала возможной только после создания Корпуса русской устной речи (http://russpeech.spbu.ru/), в котором файлы звучащей речи сопровождаются орфографической и акустико-фонетической транскрипцией. С использованием текстов транскрипции создан специализированный словарь, в котором каждому варианту акустико-фонетической транскрипции словоформы сопоставлены все возможные (встретившиеся в Корпусе) орфографические описания. На языке Python разработан алгоритм, процедура восстановления редуцированных форм в котором сводится к выбору из словаря необходимых орфографических форм для каждой «акустической» реализации словоформы, существующей в виде транскрипционной записи. Тестирование алгоритма осуществляется на материале цельных (не «расчлененных» паузами) дискурсивных единиц — межпаузальных интервалов, соответствующих клаузам и содержащих редуцированные словоформы.

Примерно треть обработанных на данный момент дискурсивных единиц интерпретируется алгоритмом однозначно просто путем обращения к описанному выше словарю, поскольку каждая из образующих клаузу редуцированных словоформ представлена в словаре единственным орфографическим вариантом. Следовательно, в таких случаях можно обойтись без морфологического анализа.

В остальных же случаях — когда одна акустическая реализация представлена в словаре несколькими орфографическими записями — приходится использовать морфологическое описание всех элементов клаузы. Для этого необходим морфологический словарь, причем не обычный, отражающий правила русской грамматики, а морфологический словарь, используемый носителем русского языка при восприятии естественной речи. Такого словаря еще не существует, поэтому приходится работать с обычным морфологическим словарем, используя в качестве промежуточного звена орфографическую запись (некоторые примеры того, какие изменения нужно при этом внести в традиционное морфологическое описание, см. [Венцов, Риехакайнен 2021]).

Мы тестируем несколько вариантов работы алгоритма. Первый из них предполагает анализ слева–направо — последовательную обработку по мере появления в речевом сигнале соответствующих лексических единиц. Предполагается, что при

[1] Санкт-Петербургский государственный университет, Российская Федерация, 199034, Санкт-Петербург, Университетская наб., 7–9

появлении в речевом потоке очередной «словоформы» в соответствующих разделах словаря активируются области со всеми возможными ее морфологиями. Появление следующей деактивирует те области, которые не соответствуют правилам сочетания. Второй подход — заполнение некоего буфера последовательно появляющимися лексическими единицами, а затем обработка, начинающаяся с анализа предиката. Еще одним вариантом является также заполнение буфера последовательно появляющимися лексическими единицами и дальнейший анализ только тех элементов, которые не имеют единственного орфографического описания, путем сопоставления их морфологического описания с ближайшими соседями. При этом любой из алгоритмов предполагает разработку и учет в программе грамматических правил (правил сочетаемости), которые бы позволили снять неоднозначность в каждом конкретном случае. Полученные на данный момент результаты не позволяют говорить о явном превосходстве какого-либо из перечисленных подходов над остальными: с одними клаузами лучше справляется один алгоритм, с другими — другие; возможно и комбинированное использование алгоритмов. При этом есть клаузы, неоднозначность в которых не удается снять ни одним из алгоритмов, поскольку требуется или более широкий контекст, или привлечение семантической информации (например, в случае необходимости выбора между союзом *но* и частицей *ну*, которые имеют совпадающий вариант произнесения [nu], или между местоимениями *то* и *это*, которые могут произноситься как [tə]).

Исследование выполнено при поддержке гранта РФФИ № 19-012-00629.

Список литературы

1. Венцов А. В., Риехакайнен Е. И. О единицах ментального лексикона носителя русского языка и их морфологическом описании // Наука СПбГУ — 2020. Сборник материалов Всероссийской конференции по естественным и гуманитарным наукам с международным участием. Санкт-Петербург, 2021. С. 1259-1260.