Автоматическая расстановка огласовок в системах распознавания арабской речи

М.Ю. Зулкарнеев,
кандидат физико-математических наук

С.А. Репалов,
кандидат физико-математических наук

С.Х. Сальман,

О.А. Свирепо,
кандидат философских наук

The paper deals with automatic diacritization in systems of automatic speech recognition for the Arabic languages. The authors give an outline of the major difficulties in Arabic the researchers usually face when they start working on the problem of automatic diacritization. This part is followed by re review of methods and approaches that have been developed. The final part is devoted to the newly developed buckWalter algorithm which allows removing incorrectly diacritized words from the dictionaries.

Арабский язык занимает пятое место в мире по количеству говорящих на нём людей. Их число по разным источникам составляет 190–290 млн человек. Этот язык чрезвычайно привлекателен для разработчиков систем автоматического распознавания речи (САРР). Между тем количество работ, посвящённых САРР для арабского языка, невелико и их показатели значительно уступают САРР для других популярных языков.

Этому факту нетрудно дать объяснение. Дело в том, что арабский язык не имеет единой унифицированной формы. Само словосочетание «арабский язык» может подразумевать:

- классический арабский язык — язык Корана и сакральных текстов;
- так называемый современный стандартизированный арабский, или литературный арабский язык (Modern Standard Arabic), понятный большинству арабоязычного населения, это язык средств массовой информации, переговоров и т.д.;
- разговорный арабский, существующий в форме диалектов.

Классический арабский язык имеет узкую сферу применения, здесь мы встретим меньший набор лексических единиц при большом разнообразии грамматических форм, многие из которых не применяется в современной разговорной речи. Литературный арабский язык используется в основном для письменной коммуникации. Разговорный вариант языка реализуется в виде десятков разновидностей диалектов. Существенное расхождение между диалектами фиксировали ещё средневековые арабские источники, датируемые X веком.

В течение последующих столетий разница в фонетическом, лексическом и грамматическом плане только усиливалась. Именно разговорная форма является, с одной стороны, самой распространённой, с другой — характеризуется самой большой вариативностью на уровне фонетики, лексики и грамматики, что выражается в разнообразии диалектов: сирийский, египетский, ливанский, марокканский и т.д. Диалекты не имеют стандартизированной письменной формы и, следовательно, не зафиксированы в форме больших собраний записанных текстов. Сложности в создании текстовой базы данных привели к тому, что большинство работ по арабскому языку выполнено на материале литературного, реже классического варианта, в то время как большая часть живой разговорной речи реализуется в форме диалектов. Эксперименты, описанные в [1], показали, что современный стандартизированный арабский язык и диалектная форма (в данном случае использовался египетский диалект) ведут себя как два совершенно разных языка. В настоящее время для исследования диалектных форм арабского языка чаще всего используются базы данных LDC (Linguistic Data Consorcium) для левантийского арабского (57,3 часа речи, в том числе 3 часа транскрибированной речи) и «Call Home» (15 часов транскрибированной речи) для египетского диалекта. Небольшой объём данных затрудняет использовать статистические подходы. Наличие большой базы данных по арабским диалектам могло бы положительно сказаться на их точности. Согласно экспериментам [2] увеличение обучающей выборки в 10 раз улучшает модель языка на 3,5%, а увеличение выборки в 8 раз улучшает акустическую модель на 5%.

Диглоссия — не единственный фактор, осложняющий жизнь разработчикам САРР. Отличительная особенность арабской письменности — отсутствие графических средств для передачи кратких гласных. Исключение составляют учебные и сакральные тексты, где для обозначения кратких гласных используются специальные значки — огласовки. В результате одна и та же графическая единица, записанная несколькими согласными буквами, может иметь несколько вариантов прочтения.

О том, какое именно значение и какую фонетическую реализацию графической формы следует выбрать, носитель языка догадывается из контекста. Разницу между чтением на европейских и арабском языках отражает поговорка: «Европеец читает, чтобы понять, араб понимает, чтобы прочитать». Разговорные варианты арабского языка используют более богатый набор фонем, чем литературный язык: в диалектах часто встречаются звуки — e, -o, однако специальных графических средств для передачи этих звуков не предусмотрено. Работа с текстами без огласовок затрудняет построение языковых и акустических моделей. Сложно обучить акустическую модель, если краткие гласные нельзя идентифицировать внутри сигнала, а также точно узнать их положение. Модель языка, обученная на неогласованных текстах, имеет меньший предсказательный потенциал, чем модель, обученная на огласованных текстах. Оба этих фактора могут отрицательно влиять на точность распознавания.

Наконец третий фактор — сложность морфологии арабского языка, затрудняющая построение модели языка. Наличие большого количества суффиксов, аффиксов, моделей словообразования даёт такое количество словоформ, что для оценки модели языка требуются текстовые базы гораздо большего объёма, чем для английского. Одно слово может представлять собой целое предложение, например burrida — «он был охлаждён». Развитая морфология также приводит к ситуации, когда система часто сталкивается со словами, которых нет в словаре. Об актуальности этой проблемы свидетельствует то, что, согласно зарубежным источникам, если использовать все сценарии словообразования для всех слов арабского языка, его словарный состав включал бы 6 x 1010 единиц [3]. Это свойство арабского языка делает

проблематичным применение получившего широкое распространение статистического подхода — слишком часто встречаются слова, которых нет в словаре. По данным [1] около 10% слов в текстах, на которых проводилось испытание, не были найдены в словаре. Примерно 50% всех этих слов были всего лишь морфологическими вариантами словарных единиц, остальные 10% — именами собственными и 40% — действительно новые слова.

Подведём итоги: в чём специфика арабского языка для систем автоматического распознавания речи? Некоторые задачи распознавания решаются проще, чем в других языках, например, создание словаря произношений, поскольку арабский даёт практически однозначное соотношение буква — звук. Наибольшую сложность при разработке высокоточной системы автоматического распознавания речи представляет текстовый материал без огласовок, сильная вариативность диалектов и морфологические трудности.

Следовательно, автоматическая расстановка огласовок — одна из важных задач при разработке системы распознания арабской речи. Автоматическая расстановка огласовок предполагает морфологический разбор слова. На данном этапе исследований существует несколько возможностей для решения этой проблемы.

Морфологический анализатор Multi-Mode Morphological Processor (MMMP), предлагаемый компанией Sakhr Software на коммерческой основе. Программа позволяет идентифицировать все возможные основы слова, часть речи, произвести морфологический разбор. Кроме того, программа может проделать и обратную работу: составить слово из его морфологических частей (основа, корень, шаблон, часть речи, аффиксы). Предполагается, что программа даёт одинаково хорошие результаты как для современного арабского языка, так и для классических текстов. В научных исследованиях редко используют ПО Sakhr ввиду его высокой стоимости и невозможности получить свободный доступ хотя бы к некоторым возможностям программы. Кроме того, компания не раскрывает информацию, касающуюся алгоритмов и принципов работы морфологического анализатора.

Ещё один вариант морфологического анализатора представлен в [4]. Создатели этой версии использовали скрытые Марковские модели (СММ), обученные на вокализованных арабских текстах. Система обучалась на текстах Корана. Полученный результат — 85% правильно расставленных огласовок, но только для текстов Корана.

В [5] для решения задачи морфологического разбора исследователи использовали конечный автомат со взвешенными состояниями, обученный на базе данных LDC (Linguistic Data Consorcium). Заявленная точность составила 93% правильно огласованных текстов.

Два последних подхода имеют некоторые недостатки. В основе статистических подходов лежит предположение о том, что текст представляет собой последовательность наблюдений (Скрытая Марковская модель), где скрытыми состояниями являются возможные символы огласовок. В обоих случаях разработчики пошли по пути создания модели и последующего обучения её на материале корпуса текстов. Следовательно, обе модели зависят от корпуса текстов, а, например, коранический корпус не отражает современных реалий арабского языка. Помимо скрытых марковских моделей, возможно применение статистических методов, как это было сделано, например в [6]. Такой подход предполагает следующую последовательность действий:

- синтаксический разбор текста: текст делится на фразы, а фразы на слова;
- морфологический анализ текста: для каждого слова предлагаются все возможные правильные варианты огласовок;
- идентификация по частям речи — выбор правильного варианта огласовки зависит, в том числе, от того, к какой части речи относится данное слово;
- применение лингвистических эвристических правил, установленных лингвистами-экспретами с целью устранить возможные оставшиеся неправильные варианты.

Исследования, проведённые Safadi et al, не были основаны на достаточном количестве текстов, поэтому полученный результат 80–90% нельзя считать абсолютно достоверным.

Как было отмечено ранее, арабский язык отличается сложной морфологией, большим количеством словообразовательных моделей, что в сочетании с малыми обучающими выборками даёт проблему большого количества слов, отсутствующих в словаре. Поэтому применение чисто статистических методов, когда неогласованное слово заменяется огласованным вариантом, наиболее часто встречающимся в тексте, даёт неудовлетворительные результаты. Значительная доля ошибок приходится на неправильно огласованные окончания.

Нами был разработан алгоритм автоматической расстановки огласовок, основанный на шаблонах слов и списках корней. Для выполнения морфологического анализа мы использовали морфологический анализатор buckWalter.

Принцип состоит в следующем. Всего в арабском языке три части речи: имя, глагол и частица. Имена и глаголы образуются на основе 3–4 буквенных корней, реже встречаются 5-буквенные корни. Таким образом, гласные в арабском языке не являются полноправными элементами корня, а передают в основном грамматическую и словообразовательную информацию, поэтому такой текст читается носителями языка легче, чем текст на индоевропейских языках с пропущенными гласными. С этой точки зрения, приблизительным аналогом консонантного письма мог бы стать текст на индоевропейском языке, использующий сокращённую запись слов без окончаний и суффиксов.

Словообразование в арабском языке происходит при помощи шаблонов. Образование однокоренных слов, относящихся к другим частям речи или другими грамматическим категориям, изображают в виде шаблонов: согласные остаются неизменными, схематически показывается чередование и/или выпадение гласных. Число таких шаблонов весьма велико, поэтому создание программы автоматического морфологического разбора для арабского языка — чрезвычайно трудоёмкое занятие, например, для создания программы MORPHO3 потребовалось 3 человеко/года.

Морфологический анализатор buckWalter, помимо того, что доступ к этой программе при условии её некоммерческого использования свободный, показал себя как эффективная основа для реализаций автоматической расстановки огласовок. На вход морфологического анализатора подаётся некоторая графическая форма слова в виде последовательности согласных букв, на выходе мы получаем все возможные правильные варианты огласовок (фактическое произношение слова). Например:

علم	@` allma
عُلِم	@` ulima
عَلَّم	@` allama
عِلْم	@` ilm
عَلَم	@` alam

Морфологический анализатор включает в себя три словаря: префиксов (a), основ (b) и суффиксов (c). Предполагается, что

- длина префикса 0–4 символа;
- основа состоит из 1–10 символов;
- суффикс может иметь 0–6 символов.

В словарях содержатся все возможные правильные варианты произнесения префиксов, основ и суффиксов соответственно. Кроме того, элементам словаря присвоено определённое грамматическое значение. Помимо словарей, морфологический анализатор также включает три файла связей ab, bc, ac, где содержится информация о всех возможных правильных сочетаниях элементов всех трёх словарей. Таким образом, графическая форма анализируется на предмет наличия в ней трёх частей — суффикса, основы, префикса и их возможных грамматических значений. Окончательный вывод о всех возможных вариантах произнесения этой последовательности букв делается на основе анализа сочетаемости. Данная модель была внедрена, при этом были дополнены словари префиксов и суффиксов морфологического анализатора buckWalter. Очевидно, что этот способ автоматической расстановки огласовок может быть использован и для автоматического распознавания записей на различных диалектах арабского языка. В этом случае все три словаря морфологического анализатора необходимо дополнить соответствующими диалектными формами.

В результате работы мы получили быстродействующий алгоритм, используемый при составлении словарей распознавания, а также возможность замены и дополнения данных словарей за короткие промежутки времени, при этом структура алгоритма на основе правил арабского языка — как литературного, так и разговорного — позволяет исключить возможность вхождения некорректных форм огласования в результирующие словари.

Литература

1. *K. Kirchhoff et. al.,* Novel Speech Recognition Models for Arabic, Final Report, JHU Summer Research Workshop, Baltimore, MD, 2002.
2. *G. Zavaliagkos, J. McDonough, A. El-Jaroudi, J. Billa, F. Richardson, K. Ma, M. Siu, H. Gish* «The BBN Byblos 1997 Large Vocabulary Conversational Speech Recognition System».
3. *Ahmad, Mohamed Attia* «A large-Scale Computational Processor of the Arabic Morphology, and Applications» A master's Thesis, Faculty of Engineering, Cairo University, Cairo, Egypt.
4. *Ya'akov Gal,* «An HMM Approach to Vowel Restoration in Arabic and Hebrew», ACL 02 Semitic Language Workshop, 2002.
5. *R. Nelken* ***и*** *S. Shieber* (Rani Nelken and Stuart M. Shieber, «Arabic Diacritization Using Weighted Finite-State Transducers», Proceedings of the 2005 ACL Workshop on Computational Approaches to Semitic Languages, pages 79–86, Ann Arbor, Michigan, June 2005.
6. *Hani Safadi, Dr. Oumayama Dakkak, Dr. Nada Ghnaim* «Computational Methods to Vocalize Arabic Texts», Proceed. of the 2006 Workshop on Internationalizing W3C's Speech Synthesis Makeup Languages.

Зулкарнеев М.Ю. —
кандидат физ.-мат. наук, старший научный сотрудник лаборатории обработки речи НИИ «Спецвузавтоматика», г. Ростов-на-Дону asni@asni.rsu.ru

Репалов С.А. —
кандидат физ.-мат. наук, заведующий лабораторией обработки речи НИИ «Спецвузавтоматика».

Сальман С.Х. —
научный сотрудник лаборатории обработки речи НИИ «Спецвузавтоматика».

Свирепо О.А. —
кандидат философских наук, старший научный сотрудник лаборатории обработки речи НИИ «Спецвузавтоматика.