РОССИЙСКАЯ АКАДЕМИЯ НАУК
ИНСТИТУТ ПРИКЛАДНОЙ МАТЕМАТИКИ имени М.В. Келдыша

Логачева В.К., Клышинский Э.С., Галактионов В.А.

Современные методы практической транскрипции

Москва
2012

Логачева В.К., Клышинский Э.С., Галактионов В.А.

Современные методы практической транскрипции

Аннотация
Статья содержит анализ существующих методов практической транскрипции. Описана общая структура системы транскрипции. Рассматриваются как статистические системы, так и системы, основанные на использовании правил. Проводится анализ основных классов статистических методов обучения: порождающих и дифференциальных.

Logacheva V.K., Klyshinsky E.S., Galaktionov V.A.

Contemporary transliteration methods

Abstract
The article contains the analysis of existing methods of machine transliteration. General structure of cross-lingual transliteration system is represented. The author considers statistic systems as well as rule-based systems, and explores two main classes of statistical methods of transliteration learning: discriminative and generative methods.

Работа выполнена при поддержке РФФИ, грант № 11-01-00793-а

Содержание

1. Введение

При переводе различных текстов (в частности, библиографических данных, адресов, географических карт) часто возникает необходимость передачи на целевой язык различных имен собственных: личных имен, географических названий и т.д. Иногда возможен перевод, если у имени есть лексическое значение, например, Easter island – остров Пасхи, Chingachgook the Great Serpent – Чингачгук Большой Змей. Такие случаи относительно редки, поэтому обычно используются другие методы.

Если перевод осуществляется между языками, пользующимися одним и тем же алфавитом, то имя может быть оставлено без перевода. Спорным случаем является использование в имени символов, отсутствующих в алфавите целевого языка. Например, многие языки пользуются различными вариантами расширенного латинского алфавита. При переводе с французского языка имя François, скорее всего, будет оставлено без перевода, хотя символ «ç» используется в алфавитах всего нескольких языков кроме французского (например, турецком и португальском).

Более актуальна проблема передачи иноязычных имен собственных из языков, использующих другую систему письма. На протяжении долгого времени их преобразование осуществлялось с помощью строгой транслитерации, то есть, сопоставлению каждой букве алфавита языка оригинала буквы алфавита целевого языка. У этого подхода есть серьезный недостаток – переведенные таким образом имена часто не сохраняют оригинального звучания в языке перевода. Однако почти до середины двадцатого века это не имело значения по нескольким причинам. Во-первых, не были развиты средства связи, передающие звуковую информацию (телефон, телевизор). Информация, получаемая из-за рубежа, была в основном текстовой, что требовало прежде всего графического, а не звукового сходства перевода имени с оригиналом. Во-вторых, звучание стало иметь значение, только когда помимо передачи иностранных имен на родной язык возникла необходимость передачи имен родного языка на иностранный (например, для оформления международных документов). В этом случае переводчик заинтересован именно в сохранении звучания (особенно, нам кажется, это касается фамильно-именных групп, так как во время пребывания за границей человек вынужден неоднократно называть свое имя, и лучше, если его графическая запись будет соответствовать звучанию).

Эти факторы не имеют прямого отношения ко многим из сфер, в которых требуется передача имен собственных между языками, но к середине двадцатого века во всех сферах распространился новый подход – передача имен собственных с сохранением их звучания. Такой подход в российской лингвистике получил название **практической транскрипции**. Этот термин впервые применён в 1935 году А. М. Сухотиным [51] и введён во всеобщее употребление А. А. Реформатским [49]. Практическую транскрипцию следует отличать от:

- фонетической транскрипции, основанной на точной передаче звучания с использованием специального фонетического алфавита;
- транслитерации, определяемой только исходным написанием;
- перевода.

В отличие от фонетической транскрипции, практическая транскрипция использует только символы алфавита языка-приёмника, а возможность введения дополнительных знаков отсутствует.

В зарубежной науке термин «transcription» является не столько лингвистическим, сколько биологическим. Процесс передачи слов некоторого языка средствами другого алфавита (не важно, сохраняется ли при этом оригинальное произношение слова) в англоязычном лингвистическом сообществе принято обозначать термином «transliteration».

Задача практической транскрипции первоначально была решена вручную путем составления систем правил транскрипции для различных пар языков. Такие системы (например, [44, 45]), использовались в качестве методических указаний для переводчиков.

Распространение ЭВМ позволило автоматизировать многие лингвистические задачи и поставило задачу автоматизации практической транскрипции. Особенно она актуальна в областях, где требуется перевод имен собственных с большого количества языков, например, при межъязыковом информационном поиске, машинном переводе, составлении электронных каталогов. В данной статье мы рассмотрим подходы к решению задачи машинной транскрипции, основные системы транскрипции, практические приложения машинной транскрипции, а также некоторые смежные задачи.

2. Общая характеристика систем машинной транскрипции

Наиболее развитые системы машинной транскрипции, как правило, имеют следующую архитектуру: они состоят из подсистемы обучения транскрипции, принимающей на вход обучающие данные (как правило, двуязычный корпус имен) и возвращающей систему правил преобразования имен с исходного языка на целевой, и подсистемы транскрипции, которая с помощью этих правил преобразовывает строки (см. рис 1).

Обучение выполняется обычно в два этапа: выравнивание имен и порождение правил транскрипции. Под выравниванием в данном случае понимается сопоставление символам, подстрокам или звукам исходного имени символов, подстрок или звуков его перевода. Затем на основе установленных соответствий составляются правила.

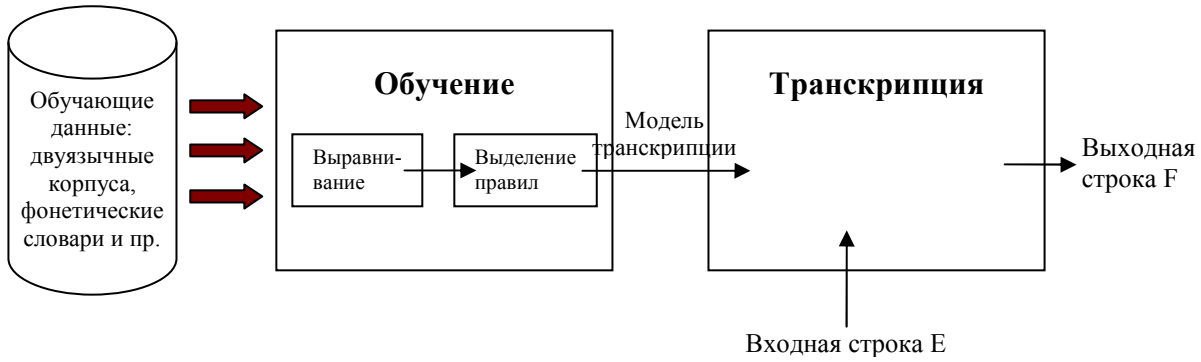

Рис. 1. Структура системы транскрипции[1]

Такой подход аналогичен методу, используемому в системах статистического машинного перевода. При обучении таких систем сначала выявляется соответствие слов оригинала словам перевода, затем формируются правила их преобразования. В некоторых методах машинной транскрипции этап выравнивания проводится с помощью статистических машинных переводчиков, например, с помощью системы GIZA++ [2, 26]. Буквы и подстроки рассматриваются системой как слова, слова – как предложения. Такой подход используется в работах [1, 13, 19, 39].

Необходимо заметить, что, в отличие от выравнивания в системах перевода, к выравниванию при транскрипции предъявляется требование сохранения порядка следования элементов (букв или звуков), тогда как при переводе сохранение порядка слов не обязательно и не всегда возможно.

Что касается метода порождения правил, наибольшей популярностью пользуются статистические методы, так как в общем случае они не требуют лингвистических знаний и дополнительной информации о языке. Почти все методы, используемые при решении задачи машинной транскрипции, можно отнести к одному из двух классов: порождающие и дифференциальные методы.

В некоторых системах используются правила, составленные вручную – например, в системе «Трансскриба» [48], системе группы Бондаренко [43], системе перевода с арабского языка на английский [5] (в этой системе используется два набора правил: соответствия между арабскими и английскими звуками и правила записи английских звуков символами латинского алфавита), а также во многих практических системах, доступных в сети Internet [47, 50, 52]. В этом случае не проводится никакого обучения.

3. Порождающие методы машинной транскрипции

Если сформулировать задачу машинной транскрипции в терминах статистических методов, то по известному входному слову E требуется найти такое $\hat{F}$ (слово на языке перевода), которое максимизирует вероятность $P(F|E)$.

[1] Рисунок взят из работы [21].

Статистические методы порождения правил транскрипции в основном опираются на формулу Байеса [6]:

$$\hat{F} = \arg\max_F \frac{P(E \mid F) \cdot P(F)}{P(E)} = \arg\max_F P(E \mid F) \cdot P(F)$$

где E – слово на языке оригинала, F – его перевод, $P(E)$ и $P(F)$ – модели языка оригинала и перевода соответственно.

Если требуется решить задачу обратной транскрипции (задачу восстановления имени на языке оригинала по имеющемуся переводу на другой язык), формула будет противоположной:

$$\hat{E} = \arg\max_E \frac{P(F \mid E) \cdot P(E)}{P(F)} = \arg\max_E P(F \mid E) \cdot P(E)$$

Можно разделить методы транскрипции на методы, использующие соответствия букв и подстрок и методы, использующие соответствия фонем. В первом случае правила транскрипции представляют собой правила перевода символов входного языка в символы выходного языка. Во втором случае правила ставят в соответствие звукам входного языка звуки выходного языка. Перед применением правил проводится преобразование имени из буквенной записи в последовательность фонем. Затем эта последовательность фонем переводится в последовательность фонем целевого языка, которая затем преобразуется в буквенную запись на целевом языке.

3.1. Методы, основанные на соответствии фонем

Первые системы машинной транскрипции использовали именно методы, основанные на соответствии фонем. Такой подход отражает суть практической транскрипции: сохранение фонетического облика слова. Схема работы подобных систем показана на рисунке 2. Преобразование производится в три этапа: оригинал имени (***E***) записывается в виде последовательности фонем входного языка (***I_E***), которая затем переводится в последовательность фонем выходного языка (***I_F***), а эта последовательность, в свою очередь, преобразуется в буквенную запись на выходном языке (***F***).

Рис.2. Схема работы системы транскрипции, основанной на соответствии фонем.

Одной из первых работ в рассматриваемой области стала система [5] транскрипции с арабского языка на английский, уже упомянутая в предыдущем разделе. Правила транскрипции в ней составлены вручную, однако она работает по той же схеме, что и остальные методы, основанные на соответствии фонем (см. рис.2).

Задача транскрипции, сформулированная в терминах фонетического сходства имен и их переводов, имеет много общего с задачей синтеза речи. В работе [11] предлагается использование методов синтеза речи для порождения правил преобразования букв к фонемам для английского и французского языков. Недостатком такого подхода является ручное составление правил.

Говоря о методах, использующих составленные вручную правила, стоит сказать об уже упоминавшейся сравнительно недавней разработке – системе «Трансскриба» [48]. В настоящее время она поддерживает 32 языка, причем возможен перевод в обе стороны для любой пары языков. Эта универсальность достигается за счет использования промежуточного фонетического представления. Правила транскрипции составляются вручную, но нет необходимости определения соответствий для каждой пары языков: достаточно определить правила преобразования букв и подстрок языка в фонетическое представление (в промежуточный фонетический алфавит по возможности включили все звуки, имеющиеся в представленных в системе языках). Точность транскрипции с помощью этой системы, как и с помощью других систем, использующих созданные вручную правила, довольно высока (около 99%).

Первой полностью автоматической (то есть, не требующей ручного составления правил ни на одном из этапов) системой транскрипции стала система Кевина Найта, впервые описанная в работе [22]. Предложенный им метод был положен в основу многих дальнейших разработок в этой области. Система Найта выполняла обратную транскрипцию с японского языка на английский (то есть, транскрипцию английских имен, записанных японской азбукой катакана, на английский язык). Транскрипция производится с помощью цепочки статистических конечных автоматов, последовательно выполняющих все этапы транскрипции (рис. 2).

Входной информацией для каждого автомата является выход предыдущего автомата. Сначала была сконструирована цепочка автоматов для осуществления прямой транслитерации: с английского языка на японский. Первый автомат делит входную строку на слова, второй преобразует слова в последовательность фонем английского языка, третий преобразует получившуюся цепочку фонем в цепочку фонем японского языка, пятый записывает японские фонемы с помощью азбуки катакана. Соответственно, строится четыре распределения вероятностей:

- $P(\boldsymbol{E})$ – модель языка;
- $P(\boldsymbol{I_E} \mid \boldsymbol{E})$ – генерация цепочки фонем по слову английского языка;
- $P(\boldsymbol{I_F} \mid \boldsymbol{I_E})$ – перевод последовательности английских фонем в последовательность японских фонем;
- $P(\boldsymbol{F} \mid \boldsymbol{I_F})$ – запись последовательности японских фонем с помощью азбуки катакана.

Требуется построить цепочку автоматов, которая максимизирует произведение

$$P(\boldsymbol{E}) \cdot P(\boldsymbol{I_E} \mid \boldsymbol{E}) \cdot P(\boldsymbol{I_F} \mid \boldsymbol{I_E}) \cdot P(\boldsymbol{F} \mid \boldsymbol{I_F})$$

Обучение конечных автоматов производится с помощью алгоритма EM (Expectation Maximization) [10, 35-37]. В качестве обучающих корпусов используются материалы американских печатных изданий, фонетический словарь университета Carnegie Mellon (Carnegie Mellon University Pronouncing Dictionary) [9], составленный вручную англо-японский корпус. Для решения задачи обратной транслитерации эта цепочка конечных автоматов применяется в обратном порядке, преобразуя запись на азбуке катакана в английское имя, которое с наибольшей вероятностью соответствует этой записи. Этот метод показал довольно высокое качество: система правильно переводила имена с японского на английский в 64% случаев, тогда как человек дал правильный ответ всего в 27% случаев.

Описанная модель, составленная из цепочки конечных автоматов, стала основой множества других методов. Джонатаном Граелом был создан инструмент, реализующий цепочку конечных автоматов, которые могут быть обучены на данных пользователя [15]. В дальнейшем он использовался во многих работах по машинной транскрипции.

Метод Найта был применен для арабского языка [3, 34]. Правда, в связи с некоторыми особенностями фонетики арабского языка и недостаточным количеством электронных корпусов качество обучения составило всего 56%.

В работе [16] предложен метод обратной транслитерации с корейского на английский. Обучение производится с помощью скрытых Марковских моделей, реализованных в виде нейронной сети с прямым ходом. Определение наиболее вероятных кандидатов в правила производится по следующей формуле:

$$
\begin{aligned}
F &= \arg\max_F P(F \mid E) \\
&= \arg\max_F P(f_1 f_2 ... f_m \mid e_1 e_2 ... e_l) \\
&= \arg\max_F P(f_1 f_2 ... f_m) \times P(e_1 e_2 ... e_l \mid f_1 f_2 ... f_m) \\
&= \arg\max_F P(I_1 I_2 ... I_m) \times P(e_1 e_2 ... e_l \mid I_1 I_2 ... I_m) \\
&\cong \arg\max_F \prod_j P(I_{t_j} \mid I_{t_{j-1}}) \times P(e_j \mid f_j)
\end{aligned}
$$

Качество транслитерации составило 56% (оценивался процент правильно переведенных слов).

Методы, основанные на соответствии фонем, имеют несколько недостатков. Во-первых, необходимость рассматривать фонетическое представление слова увеличивает количество преобразований. Каждое преобразование порождает ошибки, так что такая система менее надежна, чем система, работающая напрямую с буквенным представлением имени. Во-вторых, порождение правил преобразования во внутреннее фонетическое представление требует особых обучающих данных – фонетических словарей, таких, как Carnegie Mellon University Pronouncing Dictionary. Такие данные для большинства языков недоступны.

3.2. Методы, основанные на соответствии подстрок

Для преодоления недостатков методов, использующих соответствия фонем были предложены методы транскрипции, переводящие имя на языке

оригинала напрямую в буквенное представление на целевом языке. Эти методы не требуют специфических обучающих корпусов.

Одним из первых примеров транслитерации на основе соответствия подстрок была работа [17] 2000 года. Эта работа развивает метод транскрипции между английским и корейским [16], модель языка строится с помощью биграмм:

$$P(F) = P(f_1) \cdot \prod_{i=2}^{m} P(f_i \mid f_{i-1})$$

В этой работе качество было улучшено до 58%.

Метод n-грамм был также применен для транскрипции между английским и арабским языками [Abduljaleel], результат составил 68% правильно преобразованных имен.

Многие методы, упомянутые в предыдущей секции, были применены и для преобразования строк без промежуточного фонетического представления. В работе [4] представлен метод преобразования с помощью статистического конечного автомата, обучаемого алгоритмом EM. Использование соответствий подстрок вместо соответствий фонем улучшило результат транскрипции с помощью данного метода на 11,9%.

В работе [18], посвященной транскрипции между корейским и английским, использован новый метод выравнивания, который, в отличие от многих предыдущих методов, позволяет вводить соответствия один ко многим. Для обучения транскрипции используются деревья принятия решений. Деревья строятся с помощью алгоритма ID3 [30]. Дерево принятия решений строится для каждой буквы алфавита: 26 для английского, 46 для корейского. В качестве контекстов для каждой буквы используются три предыдущих и три следующих. Метод был опробован на обучающем корпусе, состоящем из 7000 пар имен, точность прямой транслитерации (с английского на корейский) составила 44,9%, обратной – 34,2%.

Стоит обратить внимание на работу [24] хотя бы потому что в ней затрагивается проблема транскрипции между языками, использующими латиницу, которая традиционно считается более простой задачей, чем транскрипция между разными системами письма. Работа посвящена переводу терминов, имеющих в разных языках общее происхождение, но отличающихся написанием и произношением (например, слово «способность»: *capacity* в английском языке, *capacidad* в испанском и *Kapazität* в немецком). Основой метода является вычисление обобщенной дистанции редактирования (или расстояния Левенштейна [46]) для соответствующих слов различных языков. Для каждой буквы алфавита исходного языка вычисляется наиболее вероятная операция редактирования (вставки, удаления, замены) в данном контексте. Рассматривается контекст из четырех символов:

$$\mathrm{P}(F \mid E) = \prod_{i} P(f_i \mid e_{i-2} e_{i-1} e_i e_{i+1})$$

Модель реализована в виде взвешенного конечного автомата и опробована на 1617 словах на финском, датском, нидерландском, английском, французском, немецком, итальянском, португальском и испанском языках.

Качество обучения составило 80 – 91% , качество транскрипции – 64-78%, что на 26% и 22% соответственно превосходит результат обучения с помощью простой дистанции редактирования, который был взят в качестве проверочного метода.

Таким образом, транскрипция может быть применена не только для передачи с одного языка на другой имен собственных, но и для улучшения качества машинного перевода: родственные слова (cognates) используются в системах машинного перевода для улучшения качества выравнивания при обучении на параллельных текстах [25, 27].

В работе [19] представлен алгоритм, основанный на соответствиях согласных и гласных. В работе объясняется, что для арабского, а также других языков, использующих арабский алфавит (в том числе и для персидского), фонетические методы не дают хорошего результата из-за принятого в этих языках сокращения гласных на письме. Таким образом, восстановить фонетический облик по графическому довольно сложно. Методы, преобразующие запись на языке оригинала напрямую в перевод, показывают лучшие результаты. По этой причине под гласными и согласными понимаются не звуки языка, а буквы, обозначающие эти звуки. Слова и их переводы разбиваются на подстроки вида CV, CVC, CC или VC (где C – согласная буква, V – гласная), подстроки оригинала и перевода сопоставляются друг другу, на основе этих соответствий формируется модель транскрипции. Оценка качества проводилась на составленном вручную англо-персидском корпусе из 16760 пар имен. Качество транскрипции составило 51,6%.

В следующей работе [20] этот метод был улучшен: предложен новый алгоритм выравнивания (в первой версии использовалась система Giza++). Кроме того, усовершенствован алгоритм формирования последовательностей гласных и согласных: группируются последовательности букв одного вида (только согласные или только гласные). Испытания на том же корпусе показали качество 55,3%.

В работе [32] представлен алгоритм транскрипции с английского языка на арабский. Представлены два метода: нахождение соответствий подстрок с помощью динамического программирования (алгоритм Витерби [40]) и обучение конечного автомата. Конечный автомат признан более совершенным методом, так как с его помощью может быть реализована статистическая модель языка, а также он не учитывает правила с низкой вероятностью и поддерживает правила перехода подстроки в пустое множество. Система была обучена на корпусе из 2844 пар слов, тестовое множество состояло из 300 пар, модель языка была получена отдельно на 10991 парах. Результат работы системы в проценте правильно переведенных слов не предоставлен, среднее количество ошибок в переведенных именах (средняя дистанция редактирования) – 2,01.

Предпринимались попытки использования при транскрипции личных имен дополнительной информации: языка происхождения и пола [23]. Описанный в работе метод был назван методом «семантической транслитерации», модель транскрипции определялась следующей формулой:

$$P(F \mid E) = \sum P(F \mid E, l, g) P(l, g \mid E)$$

где l – язык происхождения имени, g – пол носителя имени. В случае отсутствия какой-либо информации об имени соответствующий компонент просто изымается из модели. Метод был опробован на корпусах японских, китайских и английских имен. Лучшими результатами применения метода стали 49,4% правильно переведенных слов и 69,2% правильно переведенных символов, что хуже результатов применения к данным языкам других методов.

4. Дифференциальные методы

В отличие от порождающих методов, которые пытаются построить модель, наилучшим образом объясняющую данные, дифференциальные методы решают задачу без явного выделения правил, они не строят правильную модель, а ищут наиболее подходящую из существующих. Дифференциальные методы обычно используются для решения задач классификации или распознавания, однако были применены и к задаче машинной транскрипции.

В работе [42] транслитерация рассматривается как задача предсказания. Метод отличается от всех рассмотренных ранее: он не использует оценку условных (*P(E|F)*, *P(F|E)*) или объединенных(*P(E,F)*) вероятностных моделей языка. Помимо этого, он не требует выравнивания букв или подстрок слов оригинала и перевода: каждая буква предсказывается на основе предыдущих букв. Задача решается с помощью структурированного выходного пространства. Каждому элементу множества $H \times E$ (где E – входной алфавит, а H – история, то есть, комбинации букв, встретившиеся перед текущей) ставится в соответствие d-мерный вектор. Обучающее множество, состоящее из пар имен на языке оригинала и их переводов, преобразуется в |E| задач обучения. Каждая из этих задач – задача бинарной классификации, и для ее решения может быть использован любой метод обучения. Имея набор классификаторов, можно также оценить вероятность *P(E|H)* – вероятность появления буквы *e* после цепочки букв *h*. В описываемой работе поиск последовательности, максимизирующей *P(E|H)*, производится методом лучевого поиска.

В работе [8] дифференциальным методом решается задача обнаружения в параллельных текстах родственных слов (cognates), между которыми может быть установлено соответствие (учет подобных соответствий улучшает качество перевода). Это задача классификации методом опорных векторов, использующим вектора для подстрок, полученные подсчетом расстояния Левенштейна.

В нескольких работах был применен другой распространенный дифференциальный метод – обучение с помощью перцептрона [7, 12]. Лучшие результаты применения этого метода составили порядка 57%.

5. Смежные задачи

Ядром системы машинной транскрипции являются два метода: метод выделения правил транскрипции и метод преобразования строк по этим правилам. Однако не стоит забывать, что для возможности применения этих методов требуется подготовительная работа. Прежде всего, статистические методы обучения требуют больших обучающих корпусов. Ручное составление корпусов занимает слишком много времени, поэтому предпринимаются попытки **автоматизации сбора обучающих данных.** Для статистических моделей объем и качество корпуса имеют большое значение. Некоторые исследователи довольствуются корпусами, созданными вручную, другие используют двуязычные словари имен и терминов [22, 32, 42], получают параллельные примеры из двуязычных корпусов [33], из поисковых запросов [41], сравнивают термины из разных языков в фонетическом пространстве [14, 38]. В работе [31] описано обучение статистического конечного автомата на одноязычном корпусе.

Если отойти от теоретической проблемы обнаружения соответствий подстрок или фонем разных языков и посмотреть на задачу практической транскрипции с точки зрения конечного пользователя, возникает еще одна требующая решения задача – **определение языка происхождения имени**. При порождении модели транскрипции известен язык-источник и целевой язык, а при преобразовании некоторой произвольной строки требуется не только наличие модели транскрипции, но и знание того, какую из имеющихся моделей следует применить.

Методы распознавания языка происхождения имени делятся на две большие группы: основанные на правилах и основанные на n-граммах [28].

Методы, основанные на правилах, могут быть также разделены на две группы: методы с автоматическим построением дерева принятия решений и методы с ручным вводом характеристических правил. В случае ручного ввода специалист должен сделать вывод о том, какие конструкции не могут встречаться в данном языке, а какие, наоборот, будут для него характеристичны. Так, например, характеристичными для арабского языка или языка хинди считаются сочетания “dh” и “bh”, которые не встречаются в романских и германских языках, тогда как характерное для романских и германских языков сочетание “str” не встречается в корейском, японском и китайском. В связи с тем, что подобные правила не охватывают всех имен данного языка, метод работоспособен лишь для небольших подмножеств. В остальных случаях требуется использовать другие методы.

Метод построения деревьев принятия решений занимает промежуточное место между методом ручного ввода правил и n-граммными методами. В нем так же, как и в методе ручного ввода, берется строка произвольной длины, по которой делается вывод о принадлежности имени языку. Однако, как и в методе с ручным вводом правил, решение о принадлежности слова языку принимается детерминистически, а материала для принятия решений может оказаться не так много. Проблемой метода является тот факт, что для отнесения

слова к тому или иному языку может потребоваться анализ подстроки, совпадающей по длине с длиной самого слова, то есть анализ слова целиком.

Методы, основанные на n-граммах в чистом виде, используют следующий подход. Слово разбивается на непрерывные подстроки длины n. На обучающем корпусе проводится настройка коэффициентов, характеризующих каждую n-грамму. Обычно в качестве такой характеристики берется вероятность. Для вероятностей используются различные формулы: например, аддитивный критерий, который складывает вероятности появления n-граммы [29], или более сложные критерии, например [23]:

$$PP_c = 2^{\frac{-1}{N_c}\sum_{i=1}^{N_c}\log p(c_i|c_{i-1})}$$

6. Заключение

До начала массовой автоматизации лингвистических задач проблема практической транскрипции была решена ручным составлением систем соответствий букв или подстрок языка-источника буквам или подстрокам языка-приемника. Эти системы правил, как и сам принцип составления правил вручную, используются в некоторых системах машинной транскрипции. Однако более эффективными являются методы, извлекающие правила автоматически из двуязычного корпуса.

Первые исследователи задачи предложили переводить имя из одной буквенной записи в другую с использованием промежуточного фонетического представления. Но этот подход оказался несовершенным по нескольким причинам: не для всех языков существуют нужные для обучения корпуса, использование промежуточного представления увеличивает количество преобразований, каждое из которых создает ошибки. В связи с этим были предложены методы, ставящие в соответствие подстроки разных языков напрямую.

Все существующие методы обладают следующими недостатками. Во-первых, почти все они не универсальны, то есть, созданы для порождения правил транскрипции с английского на некоторый другой язык, причем наиболее популярными являются восточные языки: китайский, японский, корейский и др. Почти не рассматривается задача транскрипции на греческий, языки, использующие кириллический алфавит, а также между языками, использующими латиницу. Вторым недостатком является необходимость в больших обучающих корпусах. Для обучения используются статистические методы, качество работы которых сильно зависит от объема обучающих данных.

7. Список литературы

1. AbdulJaleel N., Larkey L. S. Statistical transliteration for English-Arabic cross language information retrieval. // In Conference on Information and Knowledge Management. – New Orleans, Louisiana, 2003. С. 139–146.
2. Al-Onaizan Y., Curin J., Jahr M., Knight K., Lafferty J., Melamed D., Och F. J., Purdy D., Smith N., Yarowsky D. Statistical machine translation. Tech. rep. – Johns Hopkins University, 1999.
3. Al-Onaizan Y., Knight K. Machine transliteration of names in Arabic text. // In Proceedings of the ACL workshop on Computational approaches to semitic languages. – Philadelphia, PA, 2002. С. 1–13.
4. Al-Onaizan Y., Knight K. Translating named entities using monolingual and bilingual resources. // In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. – Philadelphia, PA, 2002. С. 400–408.
5. Arbabi M., Fischthal S. M., Cheng V. C., Bart E. Algorithms for Arabic name transliteration. // *IBM Journal of research and Development 38,* 2, 1994. С. 183–194.
6. Bayes T., Price R. An Essay towards solving a Problem in the Doctrine of Chance. By the late Rev. Mr. Bayes, communicated by Mr. Price, in a letter to John Canton, M. A. and F. R. S. // Philosophical Transactions of the Royal Society of London 53, 1763. С .370 – 418.
7. Bellare K., Crammer K., Freitag D. Loss-sensitive discriminative training of machine transliteration models. // In Proceedings of the NAACL HLT Student Research Workshop and Doctoral Consortium. – Boulder, Colorado, June 2009. С. 61–65.
8. Bergsma S., Kondrak G. Alignmentbased discriminative string similarity. // In Proceedings of ACL. – June 2007. С. 656 – 663.
9. The Carnegie Mellon University Pronouncing Dictionary. [Электронный ресурс]. – Электрон. данные. – 2012. – Режим доступа: http://www.speech.cs.cmu.edu/cgi-bin/cmudict, свободный. – The CMU Pronouncing Dictionary. – Яз. англ., 01.02.2012.
10. Dempster A., Laird N., Rubin D. Maximum Likelihood from Incomplete Data via the EM Algorithm. // Journal of the Royal Statistical Society. Series B (Methodological) 39 (1) – 1977. С. 1–38.
11. Divay M., Vitale A. Algorithms for grapheme-phoneme translation for English and French: applications for database searches and speech synthesis. // Computational Linguistics 23, 4. – 1997. С. 495–523.
12. Freitag D., Khadivi S. A sequence alignment model based on the averaged perceptron. // In EMNLP-CoNLL – 2007. С. 238–247.
13. Gao W., Wong K.-F., Lam W. Phoneme-based transliteration of foreign names for OOV problem. // In *proceedings of the 1st International Joint Conference on Natural Language Processing*. Lecture Notes in Computer Science,vol. 3248. – Springer, 2004. С.110–119.

14. Goldberg Y., Elhadad M. Identification of transliterated foreign words in Hebrew script. // In Proc. of CICLing. – 2008.

15. J. Graehl. Carmel finite-state toolkit. [Электронный ресурс]. – Электрон. данные. – 1997. – Режим доступа: http://www.isi.edu/licensed-sw/carmel, свободный. – Carmel finite-state toolkit. – Яз. англ., 01.02.2012.

16. Jeong K., Myaeng S., Lee J., Choi K. Automatic identification and back-transliteration of foreign words for information retrieval. // Information Processing and Management 35, 4. – 1999. C. 523–540.

17. Kang I.-H., Kim G. English-to-Korean transliteration using multiple unbounded overlapping phoneme chunks. // In Proceedings of the 18th Conference on Computational Linguistics. – Saarbrücken, Germany, 2000. C.418–424.

18. Kang B.-J., Choi K.-S. Automatic transliteration and backtransliteration by decision tree learning. // In Conference on Language Resources and Evaluation. – Athens, Greece, 2000. C.1135–1411.

19. Karimi S., Turpin A., Scholer F. English to Persian transliteration. // In String Processing and Information Retrieval. Lecture Notes in Computer Science, vol. 4209. – Glasgow, UK, 2006. C. 255–266.

20. Karimi S. Machine transliteration of proper names between English and Persian. // Ph.D. thesis. – RMIT University, Melbourne, Australia. 2008.

21. Karimi S., Scholer F., Turpin A. Machine Transliteration Survey. // ACM Computing Surveys, Vol 43, Issue 3, Article 17. – 2011.

22. Knight K., Graehl J. Machine transliteration. // Computational Linguistics, 24(4). – 1998. C. 599–612.

23. Li H., Sim K., Kuo J.-S., Dong M. Semantic transliteration of personal names. // In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. – Prague, Czech Republic, 2007. C. 120–127.

24. Lindén K. Multilingual modeling of cross-lingual spelling variants. // Information Retrieval 9, 3. – 2005. C. 295–310.

25. Meng H., Lo W.-K., Chen B., Tang T. Generate phonetic cognates to handle name entities in English-Chinese cross-language spoken document retrieval. // In Proceedings of the IEEE workshop on Automatic Speech Recognition and Understanding. – Madonna di Campiglio, Italy, 2001. C. 311–314.

26. Och F., Ney H. A systematic comparison of various statistical alignment models. // *Computational Linguistics 29,* 1. – 2003. C. 19–51.

27. Oh J.-H., Choi K.-S. Recognizing transliteration equivalents for enriching domain-specific thesauri. // In Proceedings of the 3rd International WordNet Conference. – 2006. C.231–237.

28. Pervouchine V., Zhang M., Liu V., Li H. Improving Name Origin Recognition with Context Features and Unlabelled Data // In Proc. of COLING 2010, Beijing, pp. 972-978

29. Qu, Yan, Grefenstette. Finding ideographic representations of Japanese names written in Latin script via language identification and corpus

validation. // In Proc. 42nd ACL Annual Meeting. – 2004, Barcelona, Spain. C. 183–190,

30. Quinlan, J. R. 1986. Induction of Decision Trees. // Mach. Learn. 1, 1. – 1986. C. 81-106.

31. Ravi S., Knight K. Learning Phoneme Mappings for Transliteration without Parallel Data. // Human Language Technology Conference archive Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics. – 2009.

32. Sherif, T., Kondrak, G. Substring-based transliteration. // In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. – Prague, Czech Republic, 2007. C. 944–951.

33. Sproat R., Tao T., Zhai C. Named entity transliteration with comparable corpora. // In Proc. of ACL. – 2006.

34. Stalls B., Knight K. Translating names and technical terms in Arabic text. // In Proceedings of the COLING/ACL Workshop on Computational Approaches to Semitic Languages. – Montreal, Canada, 1998. C. 34–41.

35. Sundberg R.. Maximum likelihood theory and applications for distributions generated when observing a function of an exponential family variable. // Dissertation. – Institute for Mathematical Statistics, Stockholm University, 1971.

36. Sundberg R. Maximum likelihood theory for incomplete data from an exponential family. // Scandinavian Journal of Statistics 1 (2). – 1974. C. 49–58.

37. Sundberg R. An iterative method for solution of the likelihood equations for incomplete data from exponential families. // Communications in Statistics – Simulation and Computation 5 (1). – 1976. C. 55–64.

38. Tao T., Yoon S., Fister A., Sproat R., Zhai C. Unsupervised named entity transliteration using temporal and phonetic correlation. // In Proc. of EMNLP. – 2006.

39. Virga P., Khudanpur S. Transliteration of proper names in cross-lingual information retrieval. // In Proceedings of the ACL Workshop on Multilingual and Mixed-Language Named Entity Recognition. Sapporo, Japan, 2003. C. 57–64,

40. Viterbi A. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. // IEEE Transactions on Information Theory 13 (2). – 1967. C.260–269.

41. Wu J., Chang J. Learning to find English to Chinese transliterations on the web. // In Proc. of EMNLP/CoNLL. – 2007.

42. Zelenko D., Aone C. Discriminative methods for transliteration. // In Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. – Sydney, Australia, 2006. C.612–617.

43. Бондаренко А.В., Визильтер Ю.В., Горемычкин В.И., Клышинский Э.С., Формальный метод транскрипции иностранных имен

собственных на русский язык. // Программные продукты и системы, №1, 2010. С. 147-152.

44. Гиляревский Р.С., Старостин Б.А. Иностранные имена и названия в русском тексте. Справочник, 3 изд. / М., 1985.

45. Ермолович Д.И. Имена собственные: теория и практика межъязыковой передачи. – М.: Р.Валент, 2005.

46. Левенштейн. В. И. Двоичные коды с исправлением выпадений, вставок и замещений символов. Доклады Академий Наук СССР, 1965.

47. Лингвистический транслятор/транскриптор [Электронный ресурс]. – Электрон. данные. – 2012. – Режим доступа: http://www.lingvoconverter.com/, свободный. – Лингвистический транслятор/транскриптор. – Яз. рус., 01.02.2012.

48. Практическая транскрипция личных имен в языках народов мира, под ред. Клышинского. – М.: Наука, 2010

49. Реформатский А.А. Введение в языкознание. М., 1947.

50. Система транскрипции «Transcriptor.ru» [Электронный ресурс]. – Электрон. данные. – 2012. – Режим доступа: http://transcriptor.ru/, свободный. – Transcriptor.ru. – Яз. рус., 01.02.2012.

51. Сухотин А.М. О передаче иностранных географических названий // Вопр. географии и картографии. М., 1935. Сб. 1. С. 144-145.

52. Транскриптор студии Артемия Лебедева [Электронный ресурс]. – Электрон. данные. – 2012. – Режим доступа: http://www.artlebedev.ru/tools/transcriptor/, свободный. – Транскриптор. – Яз. рус., 01.02.2012.