Выделение русских заимствований в якутских текстах

Н. Кортегосо Виссио, В.П. Захаров

Санкт-Петербургский государственный университет

st082534@student.spbu.ru, v.zakharov@spbu.ru

Аннотация

Якутский язык включает в себя значительное количество русских заимствований. В процессе адаптации к якутскому языку русские основы могут претерпевать различные трансформации в соответствии с принципами якутского письма или сохранять исходное написание. Часто оба написания более или менее одинаково распространены для одних и тех же лексических единиц, и носители якутского языка сталкиваются с вопросом, какой вариант использовать. Правописание заимствований из русского языка актуально не только для областей языковой политики и языкового планирования: параллельно с усилиями по регламентации правильного написания ведутся исследования по установлению тенденции употребления. Для этого лексикографы должны просмотреть огромное количество письменного материала. Задачу выделения русских заимствований в якутских текстах можно сформулировать в рамках области исследования автоматической идентификации языка (language identification). Автоматическая идентификация языка (LI) относится к проблеме определения языка, на котором написан документ или его часть. В целом, LI может быть рассмотрена как задача классификации текста, то есть сопоставление документа с заранее определенным набором классов. В данной статье представлены результаты эксперимента по обучению классификатора для автоматического выделения русских заимствований в якутских текстах, сохранивших исходную орфографию. Классификатор реализован на основе модели 3-грамм.

Ключевые слова: якутский язык, русские заимствования, идентификация языка, 3-граммы, лексикография

Библиографическая ссылка: Кортегосо Виссио Н., Захаров В. П. Выделение русских заимствований в якутских текстах // Компьютерная лингвистика и вычислительные онтологии. Вып. 6 (Труды XXV Международной объединенной научной конференции «Интернет и современное общество», IMS-2022, Санкт-Петербург, 23 – 25 июня 2022 г. Сборник научных статей). — СПб.: Университет ИТМО, 2022. С. 41-54. DOI: 10.17586/2541-9781-2021-5-41-54

Введение

Якутский язык (эндоним Саха Тыла) — тюркский язык, на котором говорят около 450000 человек в Российской Федерации, большинство из которых проживают в Автономной Республике Саха (Якутия) в северо-восточной Сибири [1]. Он классифицируется как северо-восточный тюркский язык вместе с южносибирскими тюркскими языками, такими как тувинский, алтайский и хакасский.

Со времени присоединения якутов к Российскому государству в XVII веке якутский язык попал под влияние русского языка. С этого момента в лексический состав якутского языка вошло значительное количество элементов, заимствованных из русского. Такие авторы как Л.Н. Харитонов выясняют лингвистические и экстралингвистические причины, определяющие специфику функционирования и заимствований в разные периоды развития якутского языка [2].

В силу социальных и политических обстоятельств русский язык до сих пор является важным источником новых терминов и имен собственных в якутском языке.

Современный якутский пишется при помощи кириллического алфавита, который содержит все символы русского языка с добавлением 7 букв {ҕ, ҥ, h, ө, ү, дь, нь}. Гласные {е, я, ю, ё}, согласные {в, г, ж, з, ф, ц, ш, щ} и твёрдый знак используются только в заимствованиях. Поскольку якутский алфавит является надмножеством русского алфавита, термины и имена русского происхождения могут быть представлены в якутских текстах без каких-либо изменений, однако часто они также претерпевают изменения, в частности, чтобы соответствовать якутской фонетике.

Согласно Н.М. Васильевой [3] при первичном введении лексического элемента в якутский язык его основа сохраняет исходное написание, но, закрепившись в языке, обычно претерпевает различные орфографические изменения. Например, к фонетике якутского языка адаптированы такие русские заимствования, как *остуол* «стол», *куорат* «город», *бирикээс* «приказ», *сокуон* «закон», *норуот* «народ», издавна широко употреблявшиеся в разговорной речи. Такие заимствования, до сих пор не разрешенные в разговорной речи, как, например, *биисинэс* «бизнес», *сийиэс* «съезд», *эрэнгиэн* «рентген», *мусуой* «музей» обычно встречаются в текстах как в русском, так и в якутском написании. К этой категории также относится имена собственные и некоторые географические названия, например, *Дьоппуон* «Япония», *Эмиэрикэ* «Америка», *Уйбаныап* «Иванов», *Маарыйа* «Мария» и т. д. Сохраняются в русской форме лексические элементы, представляющие общественно-политические, научно-технические понятия и не поддающиеся фонетическим нормам якутского языка, например: «архитектура», «неолит», «материализм» и т. д. [3, с. 166-167]. Васильева также отмечает, что «в географических наименованиях, в мужских фамилиях в форме полного прилагательного и в названиях городов на -ск в конце пишется -ай или -эй, например, *Новай Гвинея, Охотскай муора, Пекарскай, Горькай, Курскай, Минскэй*, а также в заимствованных прилагательных, фиксируемых в русской форме, окончания передаются через -ай, -эй: *этиловай испиир, физическэй география*» [3, с. 166-167].

В словарях закрепляются наиболее распространенные варианты, и, следовательно, правописание заимствований в большей степени диктуется их употреблением. Процессы, связанные с написанием русских заимствований, являются актуальной областью лексических исследований якутского языка. С целью обнаружения тенденций употребления заимствовании в текстах лексикографы просматривают огромное количество письменного материала.

В данной статье представлен метод разработки классификатора для автоматического выделения русских заимствований, сохранивших исходное написание. Классификатор может быть применен в рамках тех исследований, где требуется выявить русские заимствования, введенные в якутский язык. Следует отметить, что понятие «заимствование» понимается здесь в широком смысле, включая лексические единицы, являющиеся устойчивой частью лексики якутского языка, а также имена собственные и иноязычные номенклатуры.

Остальная часть статьи включает краткое описание основного подхода, используемого в области компьютерной лингвистики, в рамках которого можно сформулировать данную задачу. За этим следует описание сложностей создания классификатора с учетом доступных ресурсов и предлагается метод построения классификатора и оценка его производительности. Также отдельным разделом представлены результаты выделения заимствований, сохраняющих русскую орфографию, из газеты «Саха Сирэ».

1. Предыдущие работы

Задачу выделения русских заимствований в текстах якутского языка можно сформулировать в рамках автоматической идентификации языка (language identification)

— хорошо изученного подхода в области автоматической обработки текста (АОТ). Автоматическая идентификация языка (LI) относится к проблеме определения языка, на котором написан документ или его часть. LI является ключевой частью многих процедур автоматической обработки текста, поскольку обычно необходимо определить язык исходного документа, прежде чем приступать к его дальнейшему анализу. LI применяется в основных областях АОТ, в частности, в машинном переводе, распознавании речи, в интеллектуальном анализе данных и т. д. С 1960-х гг. был разработан ряд компьютерных методов для определения используемого языка на основе правил.

В целом, LI может быть рассмотрена как задача классификации текста, то есть сопоставление документа с заранее определенным набором классов. Современные подходы к LI чаще всего используют статистические методы, которые фокусируются на распределении букв в текстах. Основная идея следующая: каждый язык имеет уникальные сочетания букв, по которым его можно идентифицировать. Это наблюдение можно формализовать с помощью модели N-грамм. N-граммы — это тип вероятностной модели для предсказания следующего элемента в последовательности N элементов. Применительно к буквам N-граммы описываются вероятностью встретить определенную букву с учетом появления перед ней N-1 других определенных букв. Например, в якутском языке употребление гласных подвергается закону гармонии гласных, по которому в рамках одной лексической единицы все гласные должны быть или только задними, или передними, то есть произнесены либо в задней, либо к передней части рта [2, с. 59]. Поскольку в русском языке нет этого ограничения, следует ожидать совершенно другого распределения гласных. Аналогичные закономерности есть и в распределении согласных.

Основная задача LI на основе N-грамм состоит в том, чтобы сгенерировать набор частотных профилей N-грамм для каждого языка, который необходимо идентифицировать. Это рассчитывается на основе корпуса. Требуется создать отдельный репрезентативный корпус для получения N-грамм профиля каждого языка. Затем для определения языка текстовой строки классификатор вычисляет вероятность ее N-грамм для каждого из языковых профилей и выбирает тот, который дает наибольшую вероятность. Более подробно тема рассмотрена в классической статье Canvar W., Trenkle J. [4].

2. Построение классификатора

Задача создания классификатора для отделения русских текстов от якутских является довольно простой. Объема оцифрованного материала на обоих языках более чем достаточно для обучения модели N-грамм. Однако построения классификатора, способного выделять заимствования, сохраняющие русское написание внутри якутских текстов, содержит некоторые специфические трудности, которые обсуждаются ниже.

2.1. Сложности при обучении моделей

Опираясь на то, что было описано в предыдущем разделе, классификатор русских заимствований требует двух профилей N-граммы: одного, моделирующего русские заимствования, которые не подвергались изменениям в соответствии с якутской фонологией, и другого, учитывающего обычные якутские звукосочетания. Получение этих профилей сопряжено с двумя проблемами. Во-первых, полагается, что якутские тексты содержат неизвестное количество русских заимствований. Наличие в них лексических единиц, сохраняющих русскую орфографию, могло внести шум в модель N-граммы. На момент исследования не были найдены текстовые источники, которые были бы современными и гарантировали отсутствие или хотя бы наличие минимального количества русизмов. Кроме того, отсутствуют данные, позволяющие приблизительно оценить наличие русских заимствований в текстах. Во-вторых, даже если набор якутских текстов был бы очищен от заимствований, остается проблема выбора текстового

материала для обучения модели. Следует отметить, что заимствования в якутских текстах утрачивают грамматические особенности, которыми они обладают в русском языке, подчиняясь правилам аффиксации якутского языка. Это значит, что заимствования, сохраняющие русскую орфографию, могут встречаться с прикрепленным якутскими аффиксами. Например: *Москваҕа* «в Москве», *космонавтар*, «космонавты». Следовательно, тексты на русском языке не являются подходящим вариантом для обучения модели русских заимствований, так как они включают словоформы с русскими флексиями, которые они теряют в якутских текстах, заменяясь аффиксами.

2.2. Исследовательский корпус

Чтобы справиться с обеими проблемами было предложено обучить обе модели на якутских текстах, применяя метод, который поясняется ниже. В качестве текстового материала использовался корпус якутского языка из проекта «Leipzig Corpora Collection Dataset», предлагающего бесплатный онлайн-доступ к 939 корпусам для 292 языков, обогащенным статистической информацией [5]. Источником данного корпуса, который доступен в версиях 2010, 2011, 2014, 2016 и 2021 годов с разными объемами, является якутская Википедия. Для обучения был выбран корпус из 2021 года, самый большой, состоящий из 100000 предложений.

Корпус был токенизирован, а токены агрегированы по частоте и отсортированы по убыванию. Также корпус был очищен, убирались те токены, которые содержали хотя бы один символ, отсутствующий в алфавите якутского. Таким образом, были удалены препинания, цифры и слова, составленные из символов других алфавитов. Также были отфильтрованы акронимы и словообразования, содержащие дефис. Все токены были преобразованы в нижний регистр. В итоге, после процесса обработки осталось 1103453 словоформы (108715 уникальных токенов).

2.3. Создание обучающих выборок

Оперируя правилами трансформации, которые проходят русские заимствования при адаптации к якутской фонетике, сформулированные Харитоновым [2, с. 59-83], был составлен фильтр для выделения тех лексических единиц, которые не подверглись описываемым ниже изменениям и, следовательно, содержат русское оригинальное написание:

— в начале якутского слова не встречается более одного согласного. Если заимствование начинается с нескольких согласных, тогда перед или между ними вставляются гласные: стакан → *ыстакаан*, класс → *кылаас*;
— в конце якутского слова не встречается более одного согласного. Исключением этого является сочетание сонорного согласного с глухим, причем сонорный всегда предшествует, например «рт» и «лт». Если заимствование оканчивается стечением согласных, тогда лишние согласные опускаются или вставляется гласный: спирт → *испиир*, морс → *муорус*;
— в середине якутского слова не встречается более двух согласных рядом. Если заимствованное слово не соответствует этому правилу, то в нем производятся соответствующие фонетические изменения: качество → *хаачыстыба*;
— в начале якутского слова никогда не встречаются согласные {ҕ, й, ҥ, р, ль}. Перед начальным «р» в русских заимствованиях вставляется гласный: рама → *араама*;
— в конце якутского слова не встречаются {б, г, ҕ, д, дь, ль, нь, h, ч}. В конце заимствований эти согласные заменяются другими: ключ → *күлүүс*.

В фильтр добавлено правило, упомянутое во введении статьи, касающееся буквенного состава заимствований:

— гласные {е, я, ю, ё}, согласные {в, г, ж, з, ф, ц, ш, щ} и твёрдый знак используются только в лексических элементах, заимствованных из русского языка.

Таким образом, фильтр состоит из 6 правил, соблюдение которых проверяется во входных лексических единицах. Если наблюдается хотя бы одно из них, то фильтр классифицирует данный лексический элемент как заимствование.

Применив фильтр на множестве всех словоформ (словоупотреблений) и на множестве уникальных словоформ, мы получили результаты, показанные в табл. 1.

Таблица 1. Обучающие выборки

	Все словоупо-требления	**Уникальные словоформы**	**500 самых частых словоформ**
Якутские словоформы	1012420 (92%)	77020 (71%)	498 (99,6%)
Русские заимствования	91033 (8%)	31695 (29%)	2 (0,4%)
Всего	1103453	108715	500

Когда фильтр применяется к уникальным словоформам, доля словоформ, классифицированных как заимствования, значительно увеличивается по отношению к общему количеству. Это понятно, если учесть, что среди 500 наиболее часто встречающихся словоформ (табл. 1, столбец 4) в корпусе встречаются только два заимствования, а именно «Россия» и «км» с рангами 228 и 274, соответственно.

В итоге получилось две обучающих выборки: одна со всеми словоупотреблениями, а другая только с уникальными словоформами.

2.4. Извлечение 3-грамм

Для обучения модели используются 3-граммы. Поскольку 3-граммы представляют собой группу из трех последовательных букв, они способны улавливать группы согласных, которые были описаны в третьем правиле фильтра в предыдущем разделе.

В процессе обучения модели на выборке с уникальными словоформами были извлечены 14262 различных 3-граммы. Из них 11857 3-грамм наблюдались в русских заимствованиях, а 6552 в якутских словоформах. В табл. 2 показаны абсолютные и относительные значения извлеченных 3-грамм.

Таблица 2. Извлеченные 3-граммы

	Количество	**Покрытие всех 3-грамм**
Из якутских словоформ	6552	45%
Из русских заимствований	11857	83%

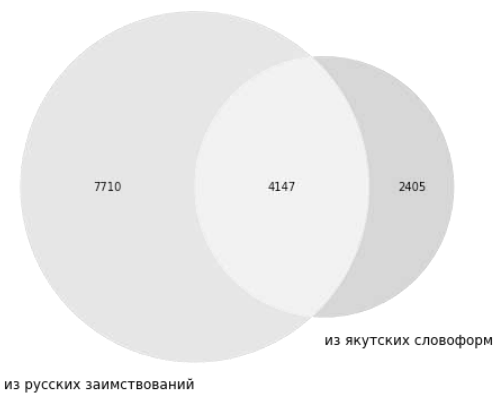

Рис. 1. Извлеченные 3-граммы

Несмотря на тот факт, что в обучающей выборке с уникальными словоформами русские заимствования представляют 29% из всех словоформ, в них содержится почти вдвое больше 3-грамм. Это показывает, что состав букв в русских заимствованиях образует больше сочетаний, чем в якутских словоформах. Из 11857 извлеченных 3-грамм 4147 являются общими для двух подмножеств (см. рис. 1).

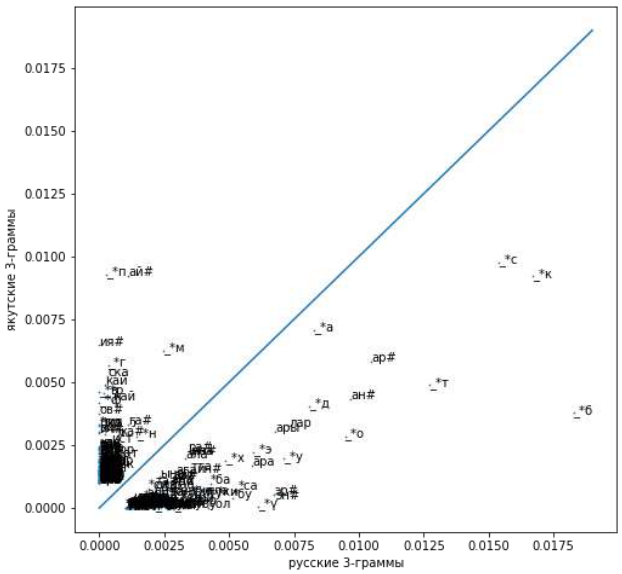

Рис. 2. Более информативные 3-граммы (все словоупотребления)

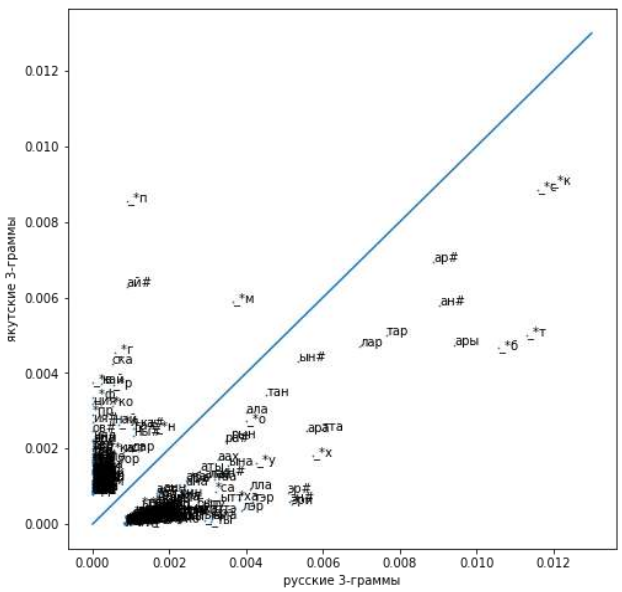

Рис. 3. Более информативные 3-граммы (уникальные словоформы)

Описанные выше данные справедливы для двух обучающих выборок, так как выборка, содержащая все словоупотребления, содержит те же самые 3-граммы, которые были выделены из выборки с уникальными словоформами. Однако, вероятности, присвоенные каждой из 3-грамм, будут разными для двух выборок.

Согласно Canvar W., Trenkle J. [4, p. 164] первые 300 N-грамм почти всегда тесно связаны с языком. Примерно с 300-го ранга профиль частоты N-грамм начинает отображать N-граммы, которые более специфичны для предмета документа, откуда был извлечены 3-граммы.

На рис. 2 и 3 показаны распределения более информативных 3-грамм для каждой обучающей выборки, то есть те 3-граммы, которые лучше отделяют русские заимствования от якутских словоформ. По оси абсцисс показаны вероятности для русских заимствований, а по оси ординат — для якутских словоформ. Из рис. 2 и 3 видно, что меняются не только значения вероятности), но и распределение 3-грамм.

3. Тестирование модели

Для проведения тестирования была составлена выборка из пары лексических элементов, которые обычно встречаются и в русском, и в якутском написании из статьи Васильевой [6]. В итоге получился список из 528 существительных без аффиксов. В табл. 3 показаны 9 примеров таких пар.

Таблица 3. Якутские-русские варианты написания

Якутский вариант написания	Русский вариант написания
остуол	стол
куорат	город
бирикээс	приказ
сокуон	закон
ачыкы	очки
киинэ	кино
тиэмэ	тема
аптамаат	автомат
биисинэс	бизнес
...	...

Чтобы установить базовый уровень, который модели 3-грамм будут способны превысить, к тестовой выборке был применен тот же фильтр, который был применен к корпусу Википедии при создания обучающей выборки. Полученные результаты представлены в табл. 4 и в виде матриц несоответствий на рис. 4, 5 и 6.

Фильтр достиг самой высокой точности: все заимствования, сохраняющие русскую орфографию, кроме одного, были классифицированы правильно. Фильтр ошибся при классификации *мэдициинэ* «медицина», потому что этот элемент включает букву «ц», что, согласно правилам, определенным в фильтре, должно встречаться только в русских заимствованиях, которые пока не претерпевали орфографические изменения. Васильева наблюдала это явление, при котором новые заимствования способствуют появлению ранее отсутствовавших звукосочетаний, которые не соответствуют якутским фонетическим нормам [6].

Таблица 4. Результаты тестирования

	Фильтр	3-граммы (все словоупотребления)	3-граммы (уникальные)
Точность	0.99	0.94	0.98
Полнота	0.84	0.96	0.97
F-мера	0.91	0.95	0.975

Рис. 4. Матрица несоответствия для фильтра

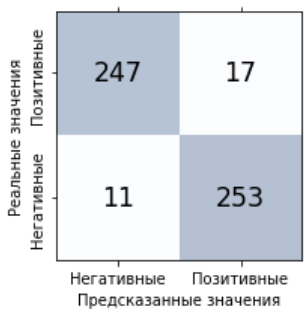

Рис. 5. Матрица несоответствия для фильтра 3-грамм (все словоупотребления)

Рис. 6. Матрица несоответствия для фильтра 3-грамм (уникальные)

Что касается полноты, то фильтр показал менее удачный результат (0.84) по сравнению с моделями 3-грамм (0.96 и 0.97). Это связано с тем, что фильтр не может идентифицировать заимствования с русским написанием, если они не показывают хотя бы одну из указанных в нем характеристик. Например, «карат», «космос», «купон», «нотариус», «пачка» и еще 38 заимствований ушли из поля зрения. Модели 3-грамм не имеют этого ограничения. Модель 3-грамм, обученная на выборке с уникальными словоформами, достигла более высокой точности на всех словоформах. Возможное объяснение более высокой производительности модели, обученной на выборке уникальных словоформ, может заключаться в том, что она уменьшает эффект 3-грамм, встречающихся в частотных словоформах. Как видно из рис. 7, наиболее частыми словами являются служебные слова, частицы, союзы, указательные и личные местоимения и т.д.

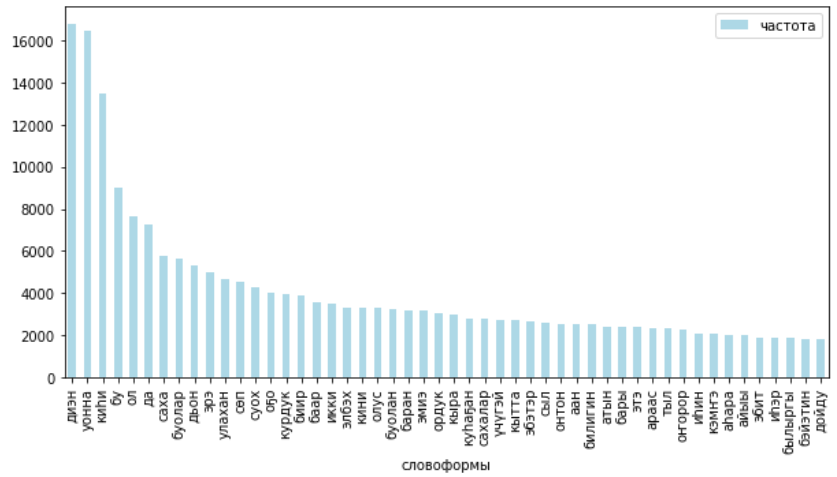

Рис. 7. 50 наиболее часто встречающихся словоформ в корпусе

С другой стороны, лексические элементы, встречающиеся с присоединенными к ним разными суффиксами, не собираются вместе (являются разными уникальными словоформами), поэтому их частота в выборке не уменьшается так резко, как таковая из служебных слов. К тому же, как было указано в разделе 3.3, русские заимствования не занимают первых позиций в частотном списке. Как показано на рис. 8, большинство из них являются уникальными.

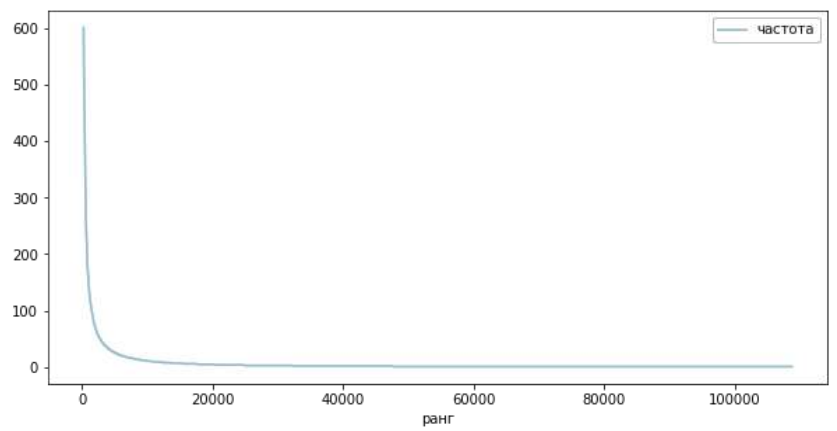

Рис. 8. Кривая Ципфа по частоте русских заимствований в обучающей выборке

Поскольку модели 3-грамм была обучена на выборке, которая ранее была получена с применением фильтра, то эти модели можно рассматривать как преобразование эвристики фильтра в статистику. Правила из фильтра становятся менее детерминированными при преобразовании в вероятностные, а также начинают фиксировать некоторые закономерности, не указанные в фильтре.

Еще одно преимущество использования статистического подхода заключается в том, что вероятности 3-граммной модели можно комбинировать с другими вероятностями, например, с априорной вероятностью нахождения заимствования, сохраняющего русскую орфографию.

На рис. 9 показано, как истинные положительные и истинные отрицательные значения для модели 3-грамм, обученной на уникальных словоформах, меняются при изменении

априорного значения от 0 до 1. Модель достигает наилучшей производительности, когда априорность установлена между 2 и 3. Повышение точности за пределами этой точки (истинно положительные) также увеличит количество истинно отрицательных результатов.

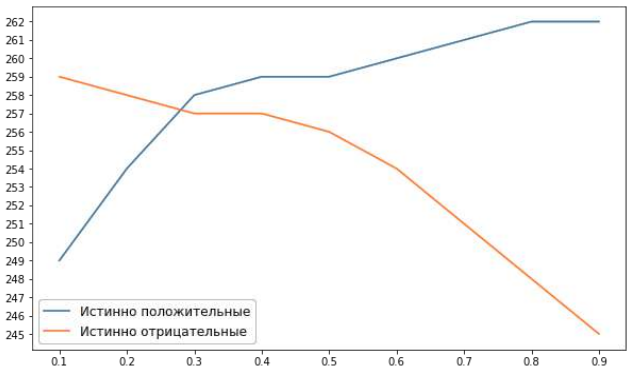

Рис. 9. Априорные значения

4. Тестирование на газетном тексте

Рассматривалась статья под названием «Арктика баайа норуот туһатыгар тахсыа дуо?» (Пойдет ли богатство Арктики на пользу народу?) из газеты «Саха Сирэ» №13 за 18. апреля 2020 г. с. 13. Саха Сирэ – это республиканская газета, публикуемая в Якутске и доступная в цифровом формате [7].

Эта статья была выбрана еще и потому, что она прямо в заголовке содержит два варианта заимствований, с которыми имеет дело текущий тест: фонетизированный вариант *норуот* «народ» и неизменяющееся русское написание «Артика». Если классификатор работает правильно, то «фонетизированное» заимствование *норуот* не должно быть выделено, Самое большое отличие от тестирования, описанного в предыдущих разделах, заключается в том, что здесь классификатор должен работать с разными частями речи и с грамматическими признаками, а не только с существительными без аффиксов.

Токенизация текста статьи проводилась по пробелам, что довольно хорошо работает для якутского языка. После устранения цифр и знаков препинания осталось 712 словоупотреблений (489 уникальных словоформ). На рис. 10 представлено распределение по частоте 60 наиболее частых словоформ в документе.

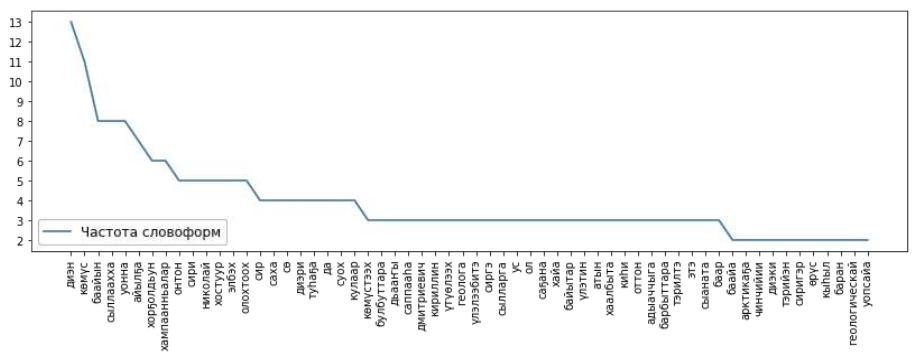

Рис. 10. Распространение словоформ в документе

Модель была обучена на выборке с уникальными словоформами, которая достигла лучших показателей чем выборка всех словоупотреблений. Классификатор выделил 90 русских заимствований в тексте (63 уникальных).

В табл. 5 показан список обнаруженных заимствованных словоформ с указанием их частотности в тексте.

Таблица 5. Список агрегированных обнаруженных заимствованных слов

mail (1), rianna789 (1), ru (1), абориген (1), арктика (1), арктикаҕа (2), атомнай (1), Валентин (1), гааска (1), Гаврил (1), геолог (1), геолога (3), геологическай (2), главсевморпуть (1), госком (1), госкомгеологияҕа (1), ГЭС (1), да (4), Дальстрой (2), Депутатскайга (2), Дмитриевич (3), Ефимов (1), империятын (1), инвестиция (1), инвестордар (2), Индигирзолото (2), Канадаҕа (1), корпоративнай (1), Куларзолото (1), кураторынан (1), м2 (1), Магадан (1), манна (1), Марианна (1), материальнай (2), металлургияҕа (1), миллиард (1), Николай (5), НКВД (1), онтон (5), протекционизм (1), регистрациялаах (1), ртуть (1), рудник (2), рудниктар (1), СГУ (1), социальнай (2), ССРС (1), стратегическай (2), субъекка (1), субъект (1), субъектарга (1), сырье (1), томпо (1), транснациональнай (1), Тыртыкова (1), фабрика (1), федеральнай (2), Цареградскай (1), Цветмет (1), экологическай (2), Юрий (1), Янзолото (1)

Четыре из них могут быть классифицированы как ложноположительные: *гааска* (газ), *манна* (здесь), *онтон* (также), *да* (союз «и») и *томпо* (выпуклый). Остальные можно считать классифицированными корректно.

ФИО, имена собственные и акронимы, содержащие чуждые якутским словоформам буквосочетания, выделяются как заимствования, так же и словоформы, написанные латиницей.

Результаты показывают, что классификатор справляется с аффиксами, например, Арктика → *Арктикаҕа* (в Арктике), рудник → *рудниктар* (рудники). Так же и в том случае, когда прикрепление аффикса изменяет корень: субъект → *субъекка*.

Таблица 6. Классификация обнаруженных заимствованных слов

Категория	Словоформы
географические места	Арктика, Арктикаҕа, Канадаҕа, Магадан
профессия	геолог, геолога, инвестордар, кураторынан
сфера труды	металлургияҕа, рудник, рудниктар, сырье, фабрика
научные или технические термины	геологическай, ртуть, экологическай, атомнай
политические или экономические термины	инвестиция, корпоративнай, миллиард, протекционизм, транснациональнай, федеральнай
аккронимы	ГЭС, м2, НКВД, СГУ, ССРС
ФИО	Валентин, Гаврил, Дмитриевич, Ефимов, Марианна, Николай, Тыртыкова, Цареградскай, Юрий
имена собственные	Главсевморпуть, Госкомгеологияҕа, Дальстрой, Депутатскайга, Индигирзолото, Куларзолото, Цветмет, Янзолото
другой алфавит	mail, rianna789, ru
другие	абориген, империятын, материальнай, социальнай, стратегическай, субъекка, субъект, субъектарга, регистрациялаах (имеющий регистрацию)

Что касается лексического значения найденных заимствований, сохраняющих русское написание, они соответствуют категориям Васильевой [3] (см. табл. 6).

Также в словоформах *геологическай*, *экологическай*, *атомнай*, *корпоративнай* и др. наблюдаются окончания -ай в прилагательных, которое было упомянуто Васильевой.

Остается проверить, как применение априорного значения может влиять на результаты. При использовании такой же априорной вероятности 0.25, которая была определена на рисунке 6 из раздела 4, классификатор выделяет 56 заимствований, то есть на 7 меньше. Отсутствуют: *ru*, *ГЭС*, *да*, *м2*, *манна*, *онтон*, *СГУ*. Два акронима утеряны, но также исчезают неправильно классифицированные местоимение *манна* и союз *онтон*.

Если учитывается количество найденных классификатором заимствований по отношению к общему количеству словоформ в документе, получается значение 0.13 (90 / 712). Используя значение 0.13 в качестве априорной вероятности, классификатор находит 54 заимствования и теряет следующие: *ru*, *гааска*, *ГЭС*, *да*, *Канадаҕа*, *м2*, *манна*, *онтон*, *СГУ*. Число ошибок уменьшается на единицу (*гааска*), но за счет потери заимствования *Канадаҕа*.

Если полнота приносится в жертву точности, то есть, при одновременном увеличении истинно положительных и ложно положительных результатов, то классификатор выделяет 67 заимствований. Новыми заимствованиями, которые были обнаружены в этом случае, являются: *Алдан* (город), *илин* (предлог «перед»), *Индигиир* (название река «Индигиирка»), *кини* (местоимение третьего лица), *онно* (нареч. «там; туда»), *онон* (союз, «поэтому; потому что»), *сотон* (глагол, «вытирать»), *сурьма*, *хостонор* (глагол, «добывается»). В этом случае верное заимствование *сурьма* выявляется ценой 8 неверных.

Заключение и дальнейшая работа

В статье представлены первые результаты эксперимента по разработке и оценке классификатора для автоматического выделения русских заимствований, сохранивших исходную орфографию в якутском языке. Был предложен метод обучения на основе 3-грамм с использованием доступных ресурсов. Результаты оказались положительными, хотя их не следует обобщать до тех пор, пока не будут проведены дополнительные эксперименты. Как и в любой другой задаче машинного обучения с учителем, производительность модели сильно зависит от обучающих данных. По мере появления более надежных обучающих данных может быть реализован более надежный подход. Код классификатора и обученные данные доступны по ссылке [8]. Нынешний классификатор может служить «металлоискателем» в лексических исследованиях. Он не гарантирует нахождения чего-то абсолютно верного, но может дать ориентир, «где копать».

К недостаткам принятого подхода можно отнести метод отбора русских заимствований, использованный при составлении обучающей выборки. Этот метод делает сильное предположение, что любая словоформа, не соответствующая правилам фильтра, является русским заимствованием. К тому же правила фильтрации могут быть пересмотрены и расширены, чтобы переобучить модели 3-грамм.

Литература

[1] Ethnologue: Languages of the World. URL: https://www.ethnologue.com/language/sah (дата обращения: 16.05.2022).

[2] Харитонов Л.Н. Современный якутский язык. Часть первая: фонетика и морфология. Научно-Исследовательский Институт языка, литературы и истории ЯАССР. Госиздат ЯАССР, Якутск, 1947.

[3] Васильева Н.М. К вопросу о правописании заимствованных слов современном якутском языке // Известия Российского государственного педагогического Университета им. А.И. Герценак. 2011. № 131. С. 166–169.

[4] Canvar W., Trenkle, J. N-Gram-Based Text Categorization // Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval. 1994. P. 161–175.

[5] Goldhahn D., Eckart Th. and Quasthoff U. Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages // Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12). 2012. P. 769–765. URL: http://www.lrec-conf.org/proceedings/lrec2012/pdf/327_Paper.pdf (дата обращения: 16.05.2022).
[6] Васильева Н. М. Об орфографической адаптации в якутском языке начальных согласных в русскоязычных заимствованиях // Научный диалог. 2018. № 5. С. 41–48.
[7] Саха Сирэ 2. апреля 2020. С. 13. URL: https://sakhamedia.ru/wp-content/uploads/2020/04/saha-sire-ot-02-aprelya-2020_compressed.pdf (дата обращения: 16.05.2022).
[8] Кортегосо Виссио Н. Классификатор языковой идентификации для извлечения русских заимствований из якутских текстов. Репозиторий на Github. URL: https://github.com/nicolascortegoso/russian_loanwords_in_yakut (дата обращения: 16.05.2022).

Identification of Russian Borrowings in Yakut Texts

N. Cortegoso Vissio, V.P. Zakharov

Saint Petersburg State University

The Yakut language includes a significant number of Russian loanwords. During the assimilation process, Russian roots may undergo transformations according to the phonetics of the Yakut language or they may retain their original spelling. Both spellings are often equally common for the same loanword, and therefore Yakut speakers are faced with the question of which variant to use.
The study of Russian loanwords in Yakut is a topic relevant not only to the areas of language policy and planning, and the efforts to standardize the spelling, but also to the research that seeks to reveal usage trends. The task of identifying Russian loanwords in Yakut texts can be carried out within the scope of the study of automatic language identification. Automatic language identification (LI) refers to the problem of determining the language in which a document or part of it is written. In general, LI can be considered as a text classification task, that is, matching a document to a set of predefined classes. This paper presents the results of an experiment on training a classifier for the automatic identification of Russian loanwords that have preserved the original spelling in Yakut texts. The classifier was trained using a 3-gram model.

Keywords: Yakut language, Russian loanwords, language identification, 3-gram model, lexicography

Reference for citation: N. Cortegoso Vissio, V. P. Zakharov. Identification of Russian borrowings in Yakut texts // Computational Linguistics and Computational Ontologies. Vol. 6 (Proceedings of the XXV International Joint Scientific Conference «Internet and Modern Society», IMS-2022, St. Petersburg, June 23-24, 2022). - St. Petersburg: ITMO University, 2022. P. 41 – 54. DOI: 10.17586/2541-9781-2021-5-41-54

Reference

[1] Ethnologue: Languages of the World. URL: https://www.ethnologue.com/language/sah (access date: 16.05.2022).
[2] Kharitonov L. N. Sovremennyy yakut·skiy yazyk. Chastʹ pervaya: fonetika i morfologiya. Nauchno-Issledovatelʹskiy Institut yazyka, literatury i istorii Yakut ASSR. Gosizdat Yakut ASSR, Yakutsk, 1947. (in Russian).
[3] Vasilʹyeva N. M. K voprosu o pravopisanii zaimstvovannykh slov sovremennom yakutskom yazyke // Izvestiya Rossiyskogo Gosudarstvennogo Pedagogicheskogo Universiteta imeni A.I. Gertsenak. 2011. № 131. P. 166–169. (in Russian).

[4] Canvar W., Trenkle, J. N-Gram-Based Text Categorization // Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval. 1994. P. 161–175.
[5] Goldhahn D., Eckart Th. and Quasthoff U. Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. // Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12). 2012. P. 769–765. URL: http://www.lrec-conf.org/proceedings/lrec2012/pdf/327_Paper.pdf (access date: 16.05.2022).
[6] Vasil′yeva N. M. Ob orfograficheskoy adaptatsii v yakutskom yazyke nachal′nykh soglasnykh v russkoyazychnykh zaimstvovaniyakh // Nauchnyy dialog. 2018. № 5. P. 41–48. (in Russian).
[7] Sakha Sire 2. April 2020. P. 13. URL: https://sakhamedia.ru/wp-content/uploads/2020/04/saha-sire-ot-02-aprelya-2020_compressed.pdf (access date: 16.05.2022). (in Russian).
[8] Cortegoso Vissio N. A language identification classifier to extract Russian loanwords from Yakut texts. Github repository. URL: https://github.com/nicolascortegoso/russian_loanwords_in_yakut (access date: 16.05.2022).