Об эвристическом методе разрешения неоднозначности при морфологическом анализе незнакомых фамилий*

Сулейманова Е. А.[1]***, Константинов К. А.***[1]
yes@helen.botik.ru
[1]Институт программных систем имени А. К. Айламазяна РАН

Статья посвящена развитию подхода к морфологическому анализу незнакомых фамилий в русскоязычном тексте, реализованного в специальном модуле системы интеллектуального анализа текста ИСИДА-Т. Идея подхода состоит в первоначальном построении заведомо избыточного множества вариантов-гипотез и последующем сокращении числа вариантов с помощью различных эвристических методов: исключение невозможных вариантов на основании дополнительных проверок правилами-фильтрами; кластеризация словоформ и фильтрация результатов внутри кластера; ранжирование вариантов по предпочтительности. Анализируются ограничения на возможности метода, вытекающие, в частности, из его детерминированной природы.

Ключевые слова: *морфологический анализ собственных имен, неоднозначность, сокращение множества гипотез, эвристические методы.*

On a heuristic approach towards ambiguity resolution in unknown surname morphological analysis*

Suleymanova E. A.[1]***, Konstantinov K. A.***[1]
[1]Program Systems Institute RAS

The paper extends the approach towards morphological analysis of unknown Russian surnames, which has been implemented as part of the ISIDA-T information extraction software. The idea is first to construct a redundant set of hypotheses and then to reduce the number of hypotheses using various heuristic techniques like ruling out impossible options with filtering rules; clustering textual forms and filtering hypotheses within clusters; preference-based ranking of options. Some limitations of the method are analysed, including those due to its deterministic nature.

Keywords: *morphological analysis of proper names, ambiguity, hypotheses reduction, heuristic methods.*

Введение

Статья посвящена развитию подхода к морфологическому анализу незнакомых фамилий в русскоязычном тексте [1], реализованному в специальном модуле системы интеллектуального анализа текста «Исида-Т» [2]. Любая система, работающая с текстом и использующая морфологический анализатор со словарем основ, сталкивается с проблемой обработки незнакомых слов — т.е. лексем, по тем или иным причинам отсутствующих в словаре. Эта проблема не решается наращиванием объема словаря, поскольку, во-первых, словообразовательные возможности языка неисчерпаемы, а во-вторых, существуют принципиально открытые категории лексики.

Работа выполнена при финансовой поддержке РФФИ, проект №13-07-00307а.

В контексте задачи извлечения информации из текстов последние представляют наибольшую важность, поскольку ФИО лиц, названия всевозможных организаций, названия географических и административных объектов служат для идентификации участников фактов.

Для обработки незнакомой (отсутствующей в словаре) лексики морфологические анализаторы обычно используют вероятностные модели предсказания, основанные на сопоставлении правых концов словоформ. Применительно к фамилиям такое морфологическое предсказание «на общих основаниях» нередко приводит к ошибкам (точные замеры нами не проводились, но некоторые показательные примеры из практики системы «Исида-Т», использующей морфологический анализатор АОТ [3], приведены в упомянутой статье [1]).

В связи с этим для системы извлечения информации был разработан специальный модуль морфологического анализа фамилий (МАФ), который работает после общего морфологического анализа.

Общее описание МАФ в системе «Исида-Т»

МАФ опирается на специально построенную модель словоизменения фамилий, которая в текущей реализации включает 59 словоизменительных классов. При построении системы классов учитывались как общие закономерности субстантивного (по типу существительного) и адъективного (по типу прилагательного) словоизменения в русском языке, так и особенности склонения фамилий. Грамматический род фамилии считается классифицирующей категорией (как род у существительного). Таким образом, фамилиям *Иванов* и *Иванова* соответствуют разные классы, так же как и мужской и женской фамилиям *Петренко*, *Обама* и т.п.

В задаче МАФ можно выделить две составляющих.

1. Лексический анализ — определить, формой какой фамилии является данная словоформа. Считаем, что фамилия однозначно определяется леммой (канонической формой) и именем словоизменительного класса. В дальнейшем фамилию как результат распознавания словоформы будем называть *вариантом фамилии*.
2. Собственно морфологический анализ — определить грамматические характеристики словоформы как формы некоторого варианта фамилии.

Неоднозначность результатов характерна как для выбора варианта фамилии, так и для определения морфологических характеристик словоформы. Основной упор при МАФ делается на разрешение лексической неоднозначности, т.е. сокращение, в идеале — до одного, числа разных вариантов фамилии для словоформы. Задача однозначного определения грамматических характеристик словоформы при этом не ставится.

Каждый результат, полученный МАФ, называется *гипотезой*. Гипотеза включает в себя вариант фамилии, вариант морфологического разбора и некоторую служебную информацию.

МАФ в системе использует фильтровый подход: для словоформ строится заведомо избыточное множество гипотез, которое затем последовательно сокращается с помощью различных методов. При таком подходе минимизируется риск упустить правильный вариант.

При построении начального множества гипотез для словоформы фамилии по возможности учитывается ее левый контекст — «этикетная» лексема (*господин*, *госпожа*, *мсье*, *леди* и т.п.), имя (известное словарю системы), возможно с отчеством. Предполагается, что значения рода и падежа всех элементов такой именной группы, включая словоформу фамилии, совпадают (число рассматривается только единственное). Учет контекста,

особенно если он в целом неомонимичный, предотвращает построение заведомо неверных гипотез уже на начальном этапе. Например, для словоформы *Проди* после омонимичного по падежу мужского имени *Романо* будут построены как гипотезы с несклоняемым вариантом мужской фамилии *Проди* для всех падежей, так и гипотеза со склоняемой мужской фамилией *Продя* в родительном падеже. Но если перед именем будет стоять неомонимичная словоформа в дательном падеже *господину*, то гипотеза для склоняемого варианта фамилии не построится (в таблице словоизменительных классов для такого варианта не найдется соответствия).

Более подробно модель словоизменения фамилий и ее использование при построении начального множества гипотез описаны в уже упомянутой статье [1].

Для дальнейшего сокращения множества гипотез применяются следующие методы.

Методы сокращения множества гипотез

Исключение некорректных вариантов с помощью правил

Проверка дополнительных условий, таких как длина леммы, число гласных букв в лемме и т.п., позволяет исключить небольшое число «запланированных» ошибочных вариантов, которые могли быть получены в результате сопоставления словоформы с таблицей словоизменительных классов (например, вариант фамилии с классом «ов/ин муж» у фамилии *Скин* — ср. *Якин*).

Фильтрация результатов путем сравнения данных из одного текста

Проанализированные к данному моменту словоформы объединяются в кластеры при помощи простого алгоритма частичного сопоставления.

Предполагается, что кластер соответствует либо одной фамилии, либо двум парным по роду фамилиям, принадлежащим в нашей модели разным классам, но «в миру» считающимся одной фамилией (как *Иванов* и *Иванова*).

Сопоставление гипотез внутри кластера с учетом рода позволяет существенно сократить множество разных вариантов фамилии для каждой словоформы. Но очевидно, что при единичном или бесконтекстном употреблении фамилии в тексте для уменьшения неоднозначности результатов требуются другие методы.

Проверка по словарю известных фамилий

Чтобы не допускать казусов в результатах работы системы извлечения информации в тех случаях, когда речь идет о фамилии какого-либо известного персонажа, а текстовых данных недостаточно для того, чтобы выбрать правильный вариант, перед применением эвристических правил предпочтения имеющиеся в кластере варианты фамилии проверяются по словарю известных фамилий. Словарь совсем небольшой, содержит порядка двух десятков фамилий (в основном иностранных, но не только — *Обама*, *Буш*, *Шойгу*).

Выбор и маркировка предпочтительных вариантов

Правила-предпочтения делятся на частные и общие. Сначала применяются частные. Если применилось хотя бы одно частное правило (условия правил сформулированы таким образом, что больше одного не может примениться), конец работы алгоритма. Иначе — переход к общим правилам предпочтения.

Частные правила-предпочтения маркируют предпочтительные варианты фамилий на основе типичных для них концовок (которые, как предполагается, маловероятны для других фамилий), например:

Если в кластере все разделы имеют совпадающую текстовую форму фамилии и она оканчивается на: `ади`, `иди`, `изи`, `оди`, `ози`, `они`, `ори`, `рти`, `жоли`, `иани`, `швили` и т.п. (реальный список значительно длиннее) и в каждом разделе есть строка с именем класса «неизм

муж» или «неизм жен», то во всём кластере строки с «неизм» пометить специальным атрибутом, остальные строки пометить как имеющие низкий приоритет.

Общие правила предпочтения основаны на довольно формальном критерии: предпочтительной считается гипотеза с максимальным *индексом совпадения*. Индекс совпадения приписывается гипотезе при построении — это число букв в ячейке таблицы словоизменительных классов, с которой успешно сопоставилась словоформа (т.е. абсолютная длина буквенного совпадения, «ответственного» за эту гипотезу). Общие правила предпочтения выполняются отдельно для вариантов фамилий мужского и женского рода.

Вывод результатов работы МАФ

К концу работы алгоритма в кластере сохраняются все гипотезы, оставшиеся после фильтрации (до проверки по словарю и применения правил предпочтения). При этом возможны следующие случаи:

— среди гипотез нет помеченных как низкоприоритетные. Это возможно в двух случаях: (1) если все неправильные гипотезы отсеялись еще на этапах фильтрации (идеальный результат) и (2) наоборот, все неправильные гипотезы, оставшиеся после фильтрации, дошли до конца работы алгоритма;

— среди гипотез есть как «предпочтительные» — помеченные специальными атрибутами (найденные по словарю или выбранные правилами предпочтения), так и низкоприоритетные.

Аннотации класса Morpho (куда в системе записываются результаты морфологического анализа) строятся по всем гипотезам, независимо от помет. Однако при построении специальных аннотаций для фамилий те аннотации Morpho, которые построены по низкоприоритетным гипотезам, из рассмотрения исключаются.

Ограничения подхода

Предел полноты

Контекст. Для того чтобы попасть в число рассматриваемых, фамилия должна хотя бы один раз встретиться в заданном контексте — либо в сопровождении «этикетной» лексемы, либо с личным именем, известным словарю личных имен системы, либо, на худой конец, с инициалом. Однократно встретившаяся одиночная фамилия, так же как и фамилия, однократно встретившаяся с неизвестным системе именем, сейчас просто игнорируется алгоритмом. Очевидно, что включить в словарь все мыслимые личные имена (особенно иностранные) невозможно. Кроме того, пополнение словаря относительно редкими именами, особенно при наличии у них омоформ среди более распространенных имен или нарицательной лексики, неизбежно приводит к увеличению «шума» (*Марин*, *Светлан*, *Танка*).

Формат. Алгоритм рассчитан на распознавание и морфологический анализ фамилий в составе относительно простой модели:

имя + (отчество)? + (среднее имя)? + («элемент фамилии»[1])? + фамилия.

При этом среднее имя должно опознаваться словарем личных имен.

Предел точности

Омоформия личных имен разного грамматического рода. Нейтрализация по роду в личном имени, с которым употреблена в тексте фамилия (*Роберт — Роберта*, *Валентин — Валентина*, *Александр — Александра*), усложняет выбор правильного варианта.

[1] Артикль или частица *фон*, *ван*, *фог*, *ле*, *де*, *ди*, *дель* и т.п.

Чем больше вхождений фамилии в тексте, тем больше шансов на правильный результат. При единичном упоминании фамилии с таким омонимичным именем алгоритм в его теперешнем виде часто оказывается не способен сделать правильный выбор.

Омонимия личных имен и прочих собственных названий. Пример регулярной ошибки: в тексте *президента Израиля Шимона Переса* алгоритм принимает за фамилию словоформу *Шимона* (*Израиль* — название государства и личное имя из словаря имен). Поскольку алгоритм фамилий работает до синтаксического анализа, предотвратить такую и подобные ей ошибки (*Рада* с фамилией *Украины*) удается только вмешательством в результаты общего морфологического анализа.

Омонимия (омоформия) личных имен и нарицательных имен. Алгоритм может «шуметь» из-за случайных совпадений слов, написанных с заглавной буквы в силу позиции, с формами личных имен из словаря. Например, любое слово с заглавной буквы после словоформы *Такая* в начале предложения может быть принято за фамилию (поскольку имя *Такай* есть в словаре имен).

Морфология составных фамилий. Фамилия из двух частей, разделенных дефисом, анализируется как одна словоформа (т.е. классифицирующая тройка для первой части не определяется), поэтому на безошибочный результат можно рассчитывать только для словоформы в канонической форме.

Зависимость от разнообразия текстовых форм. Чем меньше текстовых данных (чем более омонимичен лексический контекст и/или меньше вхождений фамилии), тем больше возрастает роль правил предпочтения. Для ограниченного числа типов фамилий с характерными «фамильными» окончаниями оказывается достаточно несложных эвристик, построенных вручную и опирающихся исключительно на буквенный состав словоформы. В огромном большинстве случаев правильный выбор может быть сделан лишь с учетом комплекса факторов разной природы, в том числе и трудно формализуемых:

— этноязыковая окрашенность имени (как минимум, русское-нерусское),
— этногеографическая привязка фамилии (может быть понятна из контекста),
— древесно-синтаксический контекст, снимающий падежную и родовую омонимию (этот уровень анализа недоступен алгоритму),
— наконец, фамилия может просто быть уже знакома из других текстов.

Примеры непростых задач для алгоритмического распознавания фамилий:

— по форме дательного падежа (*Вильгельму*) *Шойбле* предпочесть несклоняемый вариант *Шойбле* склоняемым вариантам *Шойбла* и *Шойбля* (ср. склоняемые в мужском роде фамилии *Воля*, *Сиривля*);
— по форме родительного падежа (*Марио*) *Драги* предпочесть несклоняемый вариант *Драги* склоняемому *Драга* (ср. *Брага*);
— по формам творительного падежа *Луневым*, *Кочевым* предпочесть варианты *Лунев*, *Кочев* вариантам *Луневой* и *Кочевой* (ср. *Броневой*, *Кошевой*).

Очевидно, выбор в таких случаях имеет вероятностную природу. Поэтому естественно было бы попытаться решать эту задачу статистическими методами с применением машинного обучения.

Формат. Уже упомянутое ограничение на формат приводит к потерям не только полноты, но и точности, поскольку к части «неформатных» случаев алгоритм всё-таки применяется, но с неверным результатом. Например, если в конструкции со средним име-

нем последнее не распознано словарем имен (например, *Энрике Пенья Ньето*), то оно будет разобрано как фамилия (*Энрике Пенья*).

Количественная оценка подхода

Как уже говорилось, алгоритм анализа фамилий встроен в систему извлечения информации. В настоящее время мы не располагаем отдельным инструментом для оценки количественных показателей работы алгоритма МАФ, но о качестве его работы тем не менее можно судить по результатам работы всей системы при решении задач выявления и извлечения собственных имен лиц (инструмент для оценки работы системы в целом имеется).

В задаче *выявления* оценивается факт распознавания собственного имени и правильность его границ (поскольку распознаваться должны собственные имена разной структуры и полноты). При этом правильность заполнения атрибутов собственного имени не оценивается. Можно считать, что применительно к фамилиям оценивается задача семантической классификации (является незнакомое слово фамилией или нет).

В задаче *извлечения* оценивается правильность заполнения атрибутов собственного имени. В частности, для фамилии оценивается только правильность определения леммы. Правильность определения словоизменительного класса фамилии не оценивается, т.к. для общей задачи извлечения информации это неважно, а эталонная разметка коллекции для оценки определения словоизменительного класса фамилии потребовала бы дополнительных усилий и соответствующей квалификации аннотатора.

Приведем оценку, полученную при решении задачи обнаружения и извлечения собственных имен лиц; оценка получена при очередном прогоне тестовой коллекции из 600 новостных текстов (6132 эталона) [4]:

— задача выявления собственных имен лиц: F-мера 95,50 (при точности 96,64 и полноте 94,37);
— задача извлечения собственных имен лиц: F-мера 93,58 (при точности 94,71 и полноте 92,48).

Разумеется, интересно было бы сравнить возможности предлагаемого алгоритма МАФ с другими аналогичными решениями. Однако в отсутствие общего стандарта для оценки такой задачи количественное сравнение результатов затруднительно. Нам известно описание только одного подхода к автоматическому выявлению и определению «морфологических и синтаксических характеристик» незнакомых фамилий [5]. Подход используется семантико-синтаксическим анализатором SemSin и опробован на коллекции объемом в 500 предложений. «Показано, что на новостных текстах удается опознать как фамилии, имена или инициалы до трети неизвестных слов с точностью свыше 95%». Из такой формулировки трудно понять, точность какой именно задачи оценивалась — только семантической классификации незнакомых слов или их распознавания с точностью до леммы. Кроме того, невозможно судить о показателе полноты.

Заключение

Анализ ограничений описанного подхода к морфологическому анализу незнакомых фамилий позволяет предположить, что для получения результатов приемлемого уровня на произвольной коллекции текстов требуется дальнейшее усовершенствование метода в следующих направлениях:

— использование в правилах для разрешения неоднозначности дополнительных параметров, характеризующих контекст;

— использование, наряду с «ручными» правилами, вероятностных алгоритмов, обучаемых на представительном размеченном корпусе.

Вероятностные модели вполне успешно используются для снятия частеречной омонимии при общем морфологическом анализе [6, 7, 8].

Кроме того, исследовательский интерес представляет использование результатов анализа более высоких уровней для снятия лексико-морфологической неоднозначности фамилий.

Литература

[1] *Сулейманова Е. А., Константинов К. А.* Морфологический анализ незнакомых фамилий в русскоязычном тексте // Программные продукты и системы: Международное научно-практическое приложение к международному журналу «Проблемы теории и практики управления». — 2009. — № 2. — С. 66–71.

[2] *Куршев Е. П., Кормалев Д. А., Сулейманова Е. А., Трофимов И. В.* Извлечение информации из текста в системе ИСИДА-Т // Труды XI Всероссийской научной конференции RCDL'2009, Петрозаводск: КарНЦ РАН, 2009. — С. 247–253.

[3] *Сокирко А. В.* Морфологические модули на сайте www.aot.ru // Компьютерная лингвистика и интеллектуальные технологии: Тр. Междунар. конф. Диалог'2004, М.: Наука, 2004. — С. 559–564.

[4] Коллекция «Persons-600». — `http://ai-center.botik.ru/Airec/index.php/ru/collections/27-persons-600`.

[5] *Боярский К. К., Каневский Е. А.* Автоматическое выявление фамилий в тексте // Информационные системы для научных исследований: Сборник научных статей. Труды XV Всероссийской объединенной конференции «Интернет и современное общество» (Санкт-Петербург, 10–12 октября 2012 г.), СПб, 2012. — С. 195–198.

[6] *Зеленков Ю. Г., Сегалович И. В., Титов В. А.* Вероятностная модель снятия морфологической омонимии на основе нормализующих подстановок и позиций соседних слов // Компьютерная лингвистика и интеллектуальные технологии. Тр. междунар. семинара Диалог'2005, 2005. — С. 188–197.

[7] *Сокирко А. В., Толдова С. Ю.* Сравнение эффективности двух методик снятия лексической и морфологической неоднозначности для русского языка (скрытая модель Маркова и синтаксический анализатор именных групп). — 2005. — `http://download.yandex.ru/company/grant/2005/01_Sokirko_92802.pdf`.

[8] *Sharoff S., Nivre J.* The proper place of men and machines in language technology: Processing Russian without any linguistic knowledge // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 25-29 мая 2011 г.), М.: Изд-во РГГУ, 2011. — С. 591–604.