А.В. Лаврененко (Минск, БГУ)

АЛГОРИТМ ОПРЕДЕЛЕНИЯ ИМЕН СОБСТВЕННЫХ ПРИ АВТОМАТИЧЕСКОМ АНАЛИЗЕ ТЕКСТА

Имена собственные (ИС) представляют собой достаточно информативную часть текста и поэтому нуждаются в тщательном описании. Кроме того, в системах машинного перевода определение ИС просто необходимо, поскольку ИС не подлежат переводу, а в зависимости от их типа либо остаются в тексте в неизменном виде (на языке оригинала), либо транслитерируются.

В то же время ИС создают определенную трудность при автоматической обработке текста на любом языке по следующим причинам. 1. Невозможно создать более или менее полный словарь ИС, за исключением географических и астрономических названий. 2. ИС, в отличие от обычных слов языка, по-разному оформляются (в кавычках и без них, могут состоять из одного или нескольких элементов, могут быть полным словом или аббревиатурой, как, например, инициалы). 3. ИС часто омонимичны обычным словам языка, причем не только именам существительным или прилагательным (*Анастасия Головня, Александр Зубов* (при отсутствии знака ударения ИС омонимично форме род.п. мн.ч.)*, Роберт Рождественский, медицинский центр «Здраво», журнал «Поехали!»* и т.п.). Морфологические характеристики таких ИС часто не совпадают с характеристиками обычных слов. ИС может быть несклоняемым существительным (*Республика Беларусь — в Республике Беларусь, Республикой Беларусь*) или склоняться только в некоторых условиях (в русском языке это пол обозначаемого лица: *Ивану Вечеру* и *Татьяне Вечер*). Это обстоятельство, в свою очередь, приводит к необходимости определения ИС на самых первых этапах автоматической обработки текста, чтобы на этапе синтаксического анализа не возникало проблем.

1. ИС, оформленные кавычками, можно определить на этапе графематического анализа, отделив их, во-первых, от слов, взятых в кавычки со смысловой целью, а во-вторых, от прямой речи. ИС в кавычках пишутся с заглавной буквы (хотя бы первое слово: *медаль «За отвагу на пожаре»*) и не имеют перед знаком кавычек двоеточия, а после знака кавычек — тире. Подобные наименования в кавычках на этапе графематического анализа должны быть отмечены как одна именная группа.

2. ИС, не оформленные кавычками, в середине предложения пишутся с большой буквы. Поэтому можно просмотреть весь текст, выбрать все словоформы в середине предложения с большой буквы, и составить отдельный словарь. Проверить по тексту, встречаются ли данные формы со строчной буквы, и выявить случаи омонимии. Потом проверить через этот подсловарик все слова в начале предложений. С вероятностью 90 % омонимии в реальном тексте не будет: человек старается ее избежать. Примеры типа *Воробьёв кормил воробьёв* (1), *Воробьёв кормил, воробьёв* (2), *Воробьёв кормил, Воробьёв* (3), *Воробьёв, кормил воробьёв?* (4) будут разби-

раться следующим образом: если такая фамилия встречается в тексте в середине предложения (пример 3), то пример 1 - ИС, поскольку в одном предложении омонимичные формы и вторая — со строчной буквы; пример 4 - ИС, если по тексту фамилия встречалась, и не ИС — если не встречалась; примеры 2 и 3 возможно анализировать только по контексту, поскольку они имеют повтор слов со значением смыслового усиления. Пример 2 — ответ на вопрос «Кого кормил?», пример 3 — «Кто кормил?». Но этот анализ можно проводить только тогда, когда установлено, с какими формами нарицательных имен омонимичны ИС. Это отдельный этап работы и отдельные алгоритмы.

3. ИС часто имеют при себе инициалы с точкой. Можно ожидать ошибки еще на этапе разбиения текста на предложения. Поэтому необходим предварительный анализ по всему тексту. Если это официальный документ (договор, указ, приказ и т.п.), то он должен быть подписан в самом конце текста. Две больших буквы с точками и слово с большой буквы за ними или непосредственно перед ними — это на 100% имя собственное. Чаще всего инициалы состоят из двух больших букв с точками: *А.П. Иванов*. Следует заменить пробел после второй точки каким-либо знаком или связанным пробелом и рассматривать ИС как одно слово. Варианты инициалов: *А.-Р. Супрун (Адам-Ростислав), И.-С. Бах (Иоганн-Себастьян)*. Реже встречается три буквы в инициалах.

4. Самый трудный для автоматического анализа вариант — одна заглавная буква с точкой, а после нее слово с заглавной буквы. Можно спутать с маркированным списком внутри абзаца. Чтобы исключить такую ошибку, на этапе предобработки текста просмотреть все случаи употребления одной заглавной буквы с точкой, после которой идет слово с заглавной буквы. Маркированный список обычно идет по алфавиту (от А по порядку). Ситуация, когда в произвольном тексте имена начинаются на буквы алфавита по порядку, маловероятна.

5. Одна заглавная буква с точкой может заменять целую фамилию, например: *К нам приехал господин Н.* Тогда, если точка инициала совпадает с точкой конца предложения, ситуация может быть разрешима статистикой по тексту. Если после данной буквы с точкой всегда идет одно и то же слово с заглавной буквы, то эта буква — инициал имени (тогда следующее за ней слово с заглавной буквы — фамилия, т.е. ИС), а если разные слова, то эта буква — инициал фамилии или географическое название: *Мы приехали в город Н.*

6. Инициалы часто стоят после фамилии, поэтому при обнаружении в тексте заглавных букв с последующими точками следует обратить внимание на то, с какой буквы пишется предыдущее слово. Если оно тоже с заглавной, а перед ним нет точки, то это фамилия — ИС (ввести связанный пробел и обрабатывать как единую форму). Если предыдущее слово с заглавной, а перед ним точка, то могут быть ошибки, предыдущее слово может оказаться просто началом предложения. Если предыдущее слово со строчной буквы — проверить последующее слово.

Таким образом, при отсутствии слова в словаре определить его как имя собственное можно по совокупности критериев: а) пишется в кавычках при отсутствии знаков прямой речи (да — нет); б) пишется с заглавной буквы в середине предложения (да — нет), в) имеет при себе впереди или сзади сокращения в виде заглавных букв с точкой (да - нет), г) имеет разную сочетаемость с последующим словом (для омонимов-прилагательных типа *Рождественская*), д) имеет разную сочетаемость с предыдущим словом (для омонимов-существительных типа *Воробьев, Скворец*).

В официальных документах (договорах) в начале часто дается имя, отчество и фамилия полностью, а далее по тексту — сокращение инициальное. Нужен алгоритм «восстановления» инициалов для последующей корректной морфологической обработки.