М.Н. Михайлов

ЛИПА И *ВЕНИК:* К ВОПРОСУ ОБ АННОТИРОВАНИИ ИМЕН СОБСТВЕННЫХ В КОРПУСЕ ТЕКСТОВ

Общие замечания

На первый взгляд вопрос об именах собственных кажется периферийным и сугубо техническим. Бытует мнение о некоторой лингвистической «неполноценности» имен собственных. Например, в «Частотном словаре русского языка» под ред. Л.Н. Засориной (1977)[1] таковые не представлены. Нет имен собственных и в большинстве толковых словарей. С другой стороны, не совсем понятно, чем существительное *Иван* хуже существительного *человек*. Оно является таким же русским словом со своей семантикой, морфологией и синтаксическими функциями. Даже в случае своей «неинтересности» в качестве лексемы оно может дать вполне полноценный грамматический материал. Поэтому, как нам представляется, аннотирование корпуса не может быть «выборочным». На каждом уровне должны размечаться по возможности все единицы. Представленность в корпусе разнообразной и разноплановой информации повысит эффективность работы, уменьшит количество «шума» и расширит сферу использования данных.

С технической стороны вопрос об именах собственных (ИС) также оказывается вовсе не таким простым. Написание с большой буквы вовсе не всегда означает, что слово является именем собственным. Кроме того, ИС нередко оказывается составной единицей, причем часть составляющих может писаться с малень-

[1] *Частотный словарь русского языка* / Под ред. Л.Н. Засориной. М., 1977.

кой буквы. Сравним такие ИС, как *Иван Петрович Сидоров, И.П. Сидоров, Московский государственный университет, Вторая мировая война, журнал «Новый мир»*. Дополнительную проблему представляют собой составные ИС vs. сочетания из нескольких ИС, ср. *Екатерина Дашкова – сподвижница Екатерины Второй* / *Екатерина Радищева ненавидела*. ИС образуют открытое множество, поэтому задание списков позволяет решить проблему лишь частично[1].

В настоящей статье мы попытаемся рассмотреть некоторые вопросы, возникающие при аннотировании однословных имен собственных в художественном тексте. В качестве экспериментального массива использовался небольшой корпус художественных текстов общим объемом около 2,2 млн словоупотреблений. В корпус включено 127 текстов русской художественной прозы XIX–XX вв., представлены тексты разных жанров – от рассказа-миниатюры до романа.

1. Имена и «шум»

При автоматическом аннотировании корпусов текстов нередко возникает всем знакомая техническая проблема, связанная с лемматизацией имен собственных. В списке «Грамматического словаря» А.А. Зализняка (1980)[2], использующегося при создании многих морфоанализаторов, ИС, как известно, отсутствуют. Таким образом, для обработки этих единиц приходится пополнять

[1] Подробнее см., например: *Gallippi A.F.* Learning to Recognize Names Across Languages // Proceedings of the 16th Conference on Computational Linguistics. Copenhagen, Denmark, 1996. Vol. 1. P. 424–429. URL: **http://portal.acm.org/portal.cfm**; *Wakao T., Gaizauskas R., Wilks Y.* Evaluation of an Algorithm for the Recognition and Classification of Proper Names // Proceedings of the 16th Conference on Computational Linguistics. Copenhagen, Denmark, 1996. Vol. 1. P. 418–423. **URL: http://portal.acm.org/portal.cfm**

[2] *Зализняк А.А.* Грамматический словарь русского языка. М., 1980.

список наиболее распространенными ИС, личными именами, фамилиями, географическими названиями и т.п. И даже после этого часть имен (особенно фамилии и названия населенных пунктов) не распознается вообще, поскольку их нет в списке лемматизатора. Некоторые имена интерпретируются как совпадающие с ними имена нарицательные. Особенно в этом плане проблематичны фамилии: *Пастернак* → *Пастернак/пастернак, Чехов* → *Чехов/чех, Крюков* → *Крюков/крюк* и т.д. Личные имена тоже не всегда распознаются без ошибок. Например, для имени *Корней* может быть предложена лемма *корень*, текстоформа *Тони* может быть проинтерпретирована и как английское имя *Тони*, и как форма родительного падежа от русского имени *Тоня*, и как императив от глагола *тонуть*. В литературе (например, А.П. Чехов, «В овраге») встречается имя *Липа* (уменьшительное от *Олимпиада*). Имена *Вениамин* и *Венедикт* могут сокращаться до *Веника*. Здесь же по ходу дела упомяну имя *Лампа* из детективов Д. Донцовой (уменьшительное от *Евлампия*).

В качестве имени собственного в художественном тексте может использоваться практически любое имя существительное или прилагательное (например, *Троллейбус* из романа В. Дудинцева «Белые одежды», *Гепард* из повести А. и Б. Стругацких «Парень из преисподней»). Случаи такого типа могут давать очень большое количество ненужных контекстов в конкордансах, поскольку частотность имени/прозвища одного из главных действующих лиц произведения может быть довольно высокой.

И вообще, в плане отнесения слова к именам собственным или нарицательным – как уже отмечалось выше – довольно много маргинальных случаев. Как, например, быть с названиями художественных произведений? Считать их свободными сочетаниями лексем или составными ИС? Становится ли имя нарицательное собственным, если пишется с прописной буквы для усиления (например, *Свобода*)? С другой стороны, как известно, и имена собственные могут функционировать как нарицательные (напри-

мер, *ванька* – извозчик). Такие случаи также желательно учитывать при аннотировании текстов, поскольку многие из них начинают выходить за пределы языковой игры и заслуживают отражения в словарях.

2. Заглавные буквы как основной критерий

Для распознавания имен собственных может использоваться как внутренняя структура ИС (например, последовательность «большая буква – точка – слово с большой буквы»), так и ближайший контекст (например, ключевые слова типа *город, газета* и т.п., что также позволяет выполнять и семантическое аннотирование)[1].

Разумеется, написание ИС с большой буквы – основной критерий для их распознавания в языках, в которых есть строчные и прописные буквы[2]. Главную трудность при построении алгоритма представляют слова в начале предложения или стихотворной строки, а также выделение слов путем капитализации. Поэтому алгоритм, не использующий готовых списков ИС, неизбежно будет делать ошибки. Таким образом, существующие программные продукты для поиска ИС как правило пользуются списками[3], и

---

[1] *Gallippi A.F.* Op. cit.

[2] Тем не менее, в разных языках употребление строчных и прописных букв различается, а в немецкой орфографии вообще все существительные – и собственные, и нарицательные – пишутся с большой буквы. Поэтому требуется специальная «настройка» утилит под разные языки.

[3] Stevenson M. and Gaizauskas, Robert 2000: Using Corpus-derived Name List for Named Entity Recognition. // Proceedings of the sixth conference on Applied natural language processing. Seattle, Washington, 290 – 295. (URL: http://portal.acm.org/portal.cfm). (Stevenson and Gaizauskas 2000, Gallippi 1996, Gallippi, Anthony F. 1996: Learning to Recognize Names Across Languages. // Proceedings of the 16th conference on Computational linguistics. Copenha-

лишь в некоторых случаях списки создаются в ходе «тренировки» системы[1].

Впрочем, и система, не располагающая списком ИС, имеет право на существование. Для того, чтобы в целом решить проблему заглавных букв в начале предложения (стихотворной строки) достаточно выполнить анализ вариантов написания слова в отдельном тексте и в корпусе в целом. Дополнительная трудность состоит в том, что слово может оказаться в одном тексте именем собственным, а в других – нарицательным (как, например, уже упомянутое имя *Липа* в повести Чехова «В овраге» и нарицательное существительное *липа* в остальных текстах корпуса). Написанная мной несложная утилита проверяет, встречаются ли в исследуемом тексте написания не с заглавной буквы, а затем дополнительно исследует остальные тексты корпуса. Если слово последовательно пишется с большой буквы, то случаи написания слова с большой буквы аннотируются как имена собственные.

---

gen, Denmark. – Volume 1, 424 – 429. (URL: http://portal.acm.org/portal.cfm). Wakao, Takahiro and Gaizauskas, Robert and Wilks, Yorick 1996: Evaluation of an Algorithm for the Recognition and Classification of Proper Names. // Proceedings of the 16th conference on Computational linguistics. Copenhagen, Denmark. – Volume 1, 418 – 423. (URL: http://portal.acm.org/portal.cfm).Wakao и др. 1996, Cucchiarelli, Alessandro and Luzi, Danilo and Velardi, Paola 1999: Automatic Semantic Tagging of Unknown Proper Names. // Natural Language Engineering, 5(2), 171 – 185. (URL: http://portal.acm.org/portal.cfm). Cucchiarelli и др. 1999)

[1] (Bikel и др. 1997) Bikel, Daniel M. and Miller, Scott and Schwartz, Richard and Weischedel, Ralph 1997: Nymble: a High-Performance Learning Name-finder. // Proceedings of the fifth conference on Applied natural language processing. Washington, DC, 194 – 201. (URL: http://portal.acm.org/portal.cfm).

Это позволяет программе сделать минимальное количество ошибок в тех случаях, когда в одном и том же тексте функционируют омонимичные собственные и нарицательные имена, например собственные имена *Галка, Галька, Галочка* и нарицательные *галка, галька, галочка*[1]. В случае, если слово в большинстве контекстов не в начале предложения пишется с маленькой буквы, то как ИС отмечаются только случаи написания с большой буквы не в начале предложения, во всех остальных контекстах текстоформы помечаются как нарицательные. Если же слово как правило пишется с большой буквы в середине предложения, то утилита будет отмечать все случаи написания слова с маленькой буквы как имя нарицательное, а все остальные контексты, включая начало предложения – как ИС. Такая эвристика позволит получить довольно высокое качество распознавания ИС при условии, если частотность исследуемой единицы достаточно высока. Для низкочастотных слов утилита не сможет собрать достаточно данных, и элемент случайности в полученных результатах будет достаточно высоким.

Представленность ИС в грамматическом словаре лемматизатора существенно повысит КПД программы в том плане, что появится возможность обрабатывать низкочастотные текстоформы. Слова, которые обычно являются ИС, будут анализироваться как таковые, кроме случаев, когда они пишутся с маленькой буквы. В целом количество ошибок минимально. Однако для слов, которые функционируют и как собственные, и как нарицательные (*солнце, луна, волга*), ошибок может быть несколько больше, поскольку для слова с заглавной буквы в начале предложения про-

[1] В целом такого типа случаи возможны скорее теоретически. Как отмечают Wacholder и др. (1997), автор текста – сознательно или подсознательно – избегает двусмысленных контекстов, например не начинает предложение с нарицательного существительного, если в тексте встречается омонимичное последнему имя собственное.

грамма вряд ли сможет решать однозначно, с каким случаем она имеет дело. Хуже всего обстоит дело тогда, когда ИС омонимичны именам нарицательным.

В любом случае, даже если информация об ИС представлена в словаре лемматизатора, более рациональным будет аннотирование ИС уже после завершения лемматизации, поскольку окончательное решение о статусе имени программа может принять только после получения информации о частотности написаний с большой/маленькой буквы.

3. Неравномерность распределения текстоформ внутри массива. Дополнительный критерий?

В предыдущем разделе отмечалась важность учета данных по частотности слов в корпусе. Еще одним критерием, позволяющим выделить имена собственные может стать степень равномерности распределения лексики по разным текстам корпуса и внутри текстов. В Лённгрен (1993)[1] отмечается важность этого критерия при отборе лексики для словаря: слово, имеющее высокую частотность в нескольких текстах корпуса и не встречающееся в остальных, не может включаться в словарь, даже если средняя частотность оказывается высокой.

Для нашего исследования оказывается существенным то, что степень равномерности распределения по массиву оказывается различной для слов разных грамматических классов. Более или менее стабильна только частотность служебных слов, местоимений и, в какой-то степени, наречий. Функционирование существительных и прилагательных зависит от тематики текста, поэтому даже высокочастотные слова распределяются менее равномерно, чем служебные слова. Тем не менее, распределение имен собственных оказывается еще более неравномерным.

[1] Лённгрен (1993)

Служебные слова (как и следовало ожидать) встречаются практически во всех текстах. Частотность существительных, прилагательных и глаголов может довольно сильно варьировать, однако распространенные слова имеют близкую частотность и встречаются в большей части текстов (см. рис. 1). Например, слово *лицо* не встретилось ни разу только в 12 из 127 текстов корпуса, время – в 22, глагол *идти* – в 17. Распределение ИС существенно менее стабильно. Так, имя *Иван* не встретилось ни разу в 72 текстах, и только в 8 текстах имеет относительную частоту более 5/1000 при средней частоте 1/1000. Имя *Федор* не встретилось в 106 текстах, и только в трех текстах (Дудинцев, «Белые одежды» и два рассказа Шукшина) имеет относительную частоту более 9/1000.

Распределение слов внутри текстов оказывается даже более наглядным (см. рис. 2). Имена собственные (даже имена главных героев) распределяются скачкообразно: высокая частотность перемежается полным отсутствием употреблений. Служебные слова распределяются практически равномерно, в частотности знаменательных слов хотя и наблюдаются скачки, однако ярко выраженные пики отсутствуют.

Представляется, что эта несложная эвристика может позволить усовершенствовать поиск кандидатов в имена собственные. Хотя это и не даст стопроцентных результатов, но все же несколько уменьшит объем ручной работы.

Алгоритм работы утилиты может быть, например, таким.

1. Считать очередное слово текста.

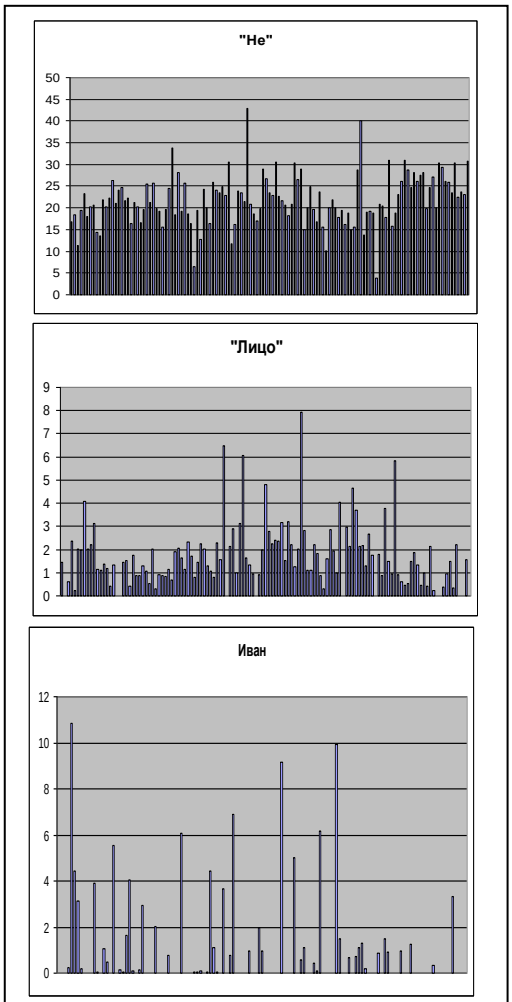


Рис. 1. Распределение слов
по текстам корпуса

2. Слово написано с большой буквы? Да: перейти к (3), нет: перейти к (1).

3. Проверить по списку ИС. Если слово в списке, перейти к (6). Если слово в списке отсутствует, перейти к (4).

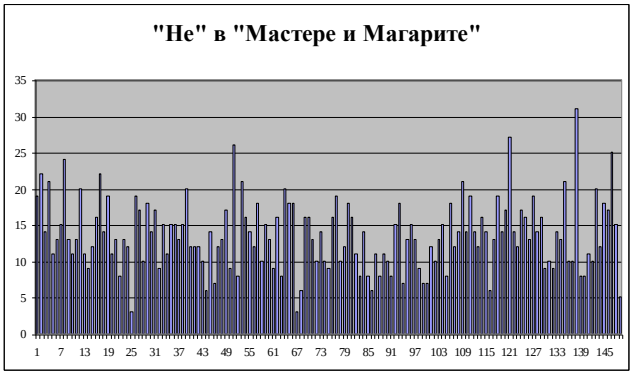
4. Проверить все другие употребления этого слова в разных текстах корпуса.

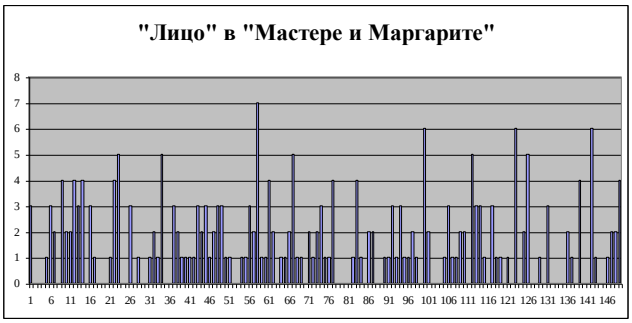
a) Если слово встретилось в корпусе только один раз не в начале предложения, отметить как ИС (например, единичные упоминания исторических деятелей, типа *Столыпин*). Перейти к (1).

b) Если слово встретилось более одного раза и всегда пишется с большой буквы, считать ИС (личные имена, которые могут отсутствовать в словаре лемматизатора, например, *Антон*, *Пульхерия*). Перейти к (1).

c) Если слово в части случаев пишется с маленькой буквы, перейти к (5).

5. Если частотность слова в некоторых текстах превышает определенную границу и распределено по тексту неравномерно (например, не встретилось в половине проверенных интервалов), считать все употребления с большой буквы в таких текстах именами собственными (включая употребления в начале предложения). Употребления с маленькой буквы считать именами нарицательными. В остальных текстах отметить это слово как ИС только при написании с большой буквы в неначальной позиции (*Липа*, *Троллейбус*). Перейти к (1).





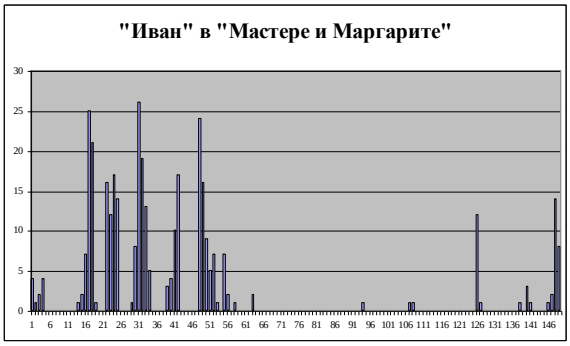


Рис. 2. Распределение слов внутри текста

6. Проверить все употребления слова в корпусе. Все случаи напи-

сания с большой буквы отметить как ИС, случаи написания с маленькой буквы отметить как нарицательные. Перейти к (1).

Такого типа алгоритм все же не позволяет корректно анализировать любые контексты. Например, в пункте 4а однократно употребленные неизвестные программе ИС не будут опознаны, если они находятся в начале предложения. В пункте 5 алгоритм может отметить имя нарицательное в начале предложения как ИС. Кроме того, если ИС омонимично служебному слову (представим себе, например, какие-нибудь имена для инопланетян типа *И* или *Для*), то алгоритм отработает некорректно.

Отметим однако, что даже человек может неправильно интерпретировать неизвестные ему имена собственные (вспомним хотя бы известную юмористическую миниатюру про студента по имени *Авас*). Поэтому задачей любого программного продукта не может быть стопроцентное распознавание, а лишь качество, приближающееся к 100%.

4. Другие проблемы и пути их решения

Некоторого снижения количества ошибок и уменьшения пропусков можно добиться, вводя формальные правила распознавания ИС. Например, можно задавать шаблоны вида «Имя – Отчество – Фамилия», «И.–О. – Фамилия» плюс описания компонентов конструкций типа «Отчество: прилагательное с финалью – *вич* или -*вна*, часто образовано от личных имен». Такие правила могут составляться вручную, а затем проверяться на экспериментальном массиве[1]. Формальные правила и учет контекста создают возможности и для семантического аннотирования имен собственных со снятием неоднозначности, что оказывается необычайно важным в информационных технологиях: системах авто-

[1] (см. Wakao и др. 1996)

матизированного перевода, поисковых системах, системах обработки сообщений на естественном языке и т.п.[1]

Создание формальных правил позволяет также аннотировать в автоматическом режиме имена собственные, состоящие из нескольких текстоформ, часть которых может писаться с маленькой буквы. Разные исследователи отмечают, что хуже всего распознаются названия организаций, несколько лучше – имена людей и лучше всего – топонимы[2].

Довольно важным, в том числе и для аннотирования художественных текстов, является аннотирование имен, имеющих одного и того же референта, например *Лев Толстой, Лев Николаевич Толстой, Л.Н. Толстой, граф Толстой* и т.п. Этот вопрос также в принципе решается путем поиска совпадений части составляющих имени[3].

Таким образом, проблема аннотирования имен собственных в корпусе текстов может быть решена с разной степенью глубины. Требуется ли семантическое аннотирование кроме грамматического, нужно ли искать составные ИС и как аннотировать их компоненты? Все эти вопросы решают разработчики корпуса текстов. Ясно только одно: корпус текстов, в котором вопрос об аннотировании ИС игнорируется, проигрывает очень сильно.

Литература

Зализняк А.А. 1980: Грамматический словарь русского языка. Москва: «Русский язык».

[1] (Wacholder и др. 1997).

[2] (см. Gallippi 1996, Stevenson and Gaizauskas 2000)

[3] (Wacholder и др. 1997).

Лённгрен Леннарт (ред.) 1993: Частотный словарь современного русского языка. Uppsala, Acta Universitatis Upsaliensis, Studia Slavica Upsaliensia 32.

Bikel, Daniel M. and Miller, Scott and Schwartz, Richard and Weischedel, Ralph 1997: Nymble: a High-Performance Learning Name-finder. // Proceedings of the fifth conference on Applied natural language processing. Washington, DC, 194 – 201. (URL: http://portal.acm.org/portal.cfm).

Cucchiarelli, Alessandro and Luzi, Danilo and Velardi, Paola 1999: Automatic Semantic Tagging of Unknown Proper Names. // Natural Language Engineering, 5(2), 171 – 185. (URL: http://portal.acm.org/portal.cfm).

Gallippi, Anthony F. 1996: Learning to Recognize Names Across Languages. // Proceedings of the 16th conference on Computational linguistics. Copenhagen, Denmark. – Volume 1, 424 – 429. (URL: http://portal.acm.org/portal.cfm).

Stevenson, Mark and Gaizauskas, Robert 2000: Using Corpus-derived Name List for Named Entity Recognition. // Proceedings of the sixth conference on Applied natural language processing. Seattle, Washington, 290 – 295. (URL: http://portal.acm.org/portal.cfm).

Wachholder, Nina and Ravin, Yael and Choi, Misook 1997: Disambiguation of Proper Names in Text. // Proceedings of the fifth conference on Applied natural language processing. Washington, DC, 202 – 208. (URL: http://portal.acm.org/portal.cfm).

Wakao, Takahiro and Gaizauskas, Robert and Wilks, Yorick 1996: Evaluation of an Algorithm for the Recognition and Classification of Proper Names. // Proceedings of the 16th conference on Computational linguistics. Copenhagen, Denmark. – Volume 1, 418 – 423. (URL: http://portal.acm.org/portal.cfm).