Некоторые сложности автоматизированной лемматизации несловарных словоформ

Some difficulties in automated lemmatization of word forms not contained in the dictionary

Клышинский Э. С. (klyshinsky@mail.ru)

Институт прикладной математики им. М. В. Келдыша РАН

В статье рассмотрены результаты машинного эксперимента по лемматизации несловарных словоформ. Рассматриваются некоторые сложности, возникающие в процессе лемматизации. В заключении делается вывод о невозможности на данный момент полностью автоматизировать процесс включения новых слов в морфологический словарь.

1. Введение

На данный момент создано большое количество компьютерных программ словарной морфологии. Некоторые из них сочетают сразу несколько функций (см., например, [1, 2]). Но даже в этом случае в первую очередь в словарь должны быть занесены лексемы, которым в дальнейшем будет привязываться другая информация.

В связи с активным наполнением словарей последнее время начали активно развиваться методы автоматизированной лемматизации словоформ [3, 4]. Каждая серьезная система морфологического анализа обладает возможностью порождения гипотез относительно нормальной формы и набора параметров незнакомого слова. Однако на практике данные методы не позволяют полностью автоматизировать процесс лемматизации несловарных словоформ, так как количество порождаемых гипотез зачастую слишком велико.

В данной работе мы рассмотрим результаты машинного эксперимента по автоматической лемматизации несловарных (то есть отсутствующих в данном словаре) словоформ. Также мы попытаемся рассмотреть некоторые трудности, которые мешают успешной автоматизации данного процесса.

2. Метод исследования

В данной работе была поставлена цель оценить эффективность использования системы автоматической лемматизации для задач наполнения морфологического словаря. Исследования проводились на подсистеме морфологического анализа системы «Кросслятор 2.0» [2]. Объем словаря — почти 160 тыс. словооснов. Для тестирования использовалось несколько корпусов текстов, представленных в открытом доступе в сети Интернет:

1. фрагмент Национального корпуса русского языка, объем более 900 тыс. словоформ по которым было порождено более 36 тыс. несловарных лексем и 1,7 млн. словарных лексем;
2. информация с сайта bash.org.ru, объем более 620 тыс. словоформ по которым было порождено более 65 тыс. несловарных лексем и более миллиона словарных лексем;
3. текущие новости с портала lenta.ru за период с марта 2005 по декабрь 2008, объем более 24,4 млн. словоформ по которым было порождено более 1,1 млн. несловарных лексем и около 46,6 млн. словарных лексем;
4. текущие новости с портала rbc.ru за период с января 2003 по декабрь 2008, объем более 17,3 млн. словоформ по которым было порождено около 2,1 млн. несловарных лексем и более 32,8 млн. словарных лексем;
5. литературный портал lib.ru, объем более 688,5 млн. словоформ по которым было порождено около 37,1 млн. несловарных лексем и более 1,3 млрд. словарных лексем;
6. материалы конференции Диалог с 2003 по 2008 год, объем около 730 тыс. словоформ по которым было порождено более 41 тыс. несловарных лексем и около 1,3 млн. словарных лексем;

Выделение словоформ проводилось простейшим путем, то есть не учитывалась возможность

разрыва слова знаком препинания, например, дефисом. В связи с этим такие слова, как «как-то», «кто-нибудь», «Аддис-Абеба» и другие, рассматривались по частям и, как следствие, могли попасть в список несловарных, хотя слово целиком находится в словаре. Однако фрагменты общеупотребительных слов (например, «нибудь») и фрагменты, являющиеся самостоятельными словами («как», «то»), к началу эксперимента были представлены в словаре.

Омонимия никак не разрешалась, то есть одна словоформа могла участвовать в полученном результате несколько раз. Однако, как это видно из приведенных выше цифр, средний уровень омонимии не превышал двух.

В ходе выдвижения гипотез о лемме несловарной словоформы проводилась активная фильтрация полученных результатов. При этом использовалось несколько сильных, но интуитивно верных положений.

1. Гипотезы, порожденные на основе редковстречающихся парадигм, в рассмотрение не брались. Под редковстречающейся понималась парадигма, количество слов которой не превышало заданного порога. При среднем количестве слов в одной парадигме около 50, брались два значения этого порога: 5 и 10 слов.
2. Для словарных слов, принадлежащей одной парадигме, определялся список букв, заканчивающих их псевдоосновы. В случае, если для словоформы выдвигалась гипотеза о ее принадлежности к данной парадигме, и если при этом ее псевдооснова не оканчивалась ни на одну из полученных букв, то такая гипотеза отвергалась.
3. Отсеивались гипотезы, образованные от словоформы, встретившейся единственный раз в исследуемом корпусе и являющиеся единственной словоформой, использованной в данной парадигме. Предполагалось, что подобная словоформа скорее всего написана с ошибкой. Исключение делалось для парадигм не изменяющихся слов (то есть содержащих единственную позицию в парадигме).
4. Считалось, что псевдооснова несловарных словоформ, объединяемых в рамках одной парадигмы, должна содержать хотя бы один символ.

Применение этих положений позволило сократить количество анализируемой информации до приемлемого уровня, позволившего перейти к процессу кластеризации словоформ [5].

При кластеризации объединялись все словоформы с одинаковой псевдоосновой и образованные по единой парадигме. При этом считалось, что одна и та же словоформа может быть омонимичной, и ей разрешалось входить в несколько гипотез. Подобный подход позволил несколько улучшить результат, оставив в нем правильные варианты. При этом, однако, количество оставляемых гипотез заметно возросло.

После кластеризации проводилось отсеивание полученных лемм по критерию максимальной встречаемости словоформ, вошедших в лемму. То есть для каждого слова определялось сколько раз оно встретилось в тексте. Далее эти значения суммировались по парадигмам и оставлялись лишь парадигмы с максимальной сумой.

В итоге генерировалось два списка лемм (словарных и несловарных) с привязанными к ним словоформами. Для каждого списка генерировалась статистика заполнения парадигм, то есть количество гипотез в зависимости от процента позиций, занятых в парадигме. Для данной статистики анализировалось количество парадигм, заполненных более чем на половину. Предполагалось, что подобные парадигмы дают приемлимый результат с точки зрения количества порождаемых гипотез. На самом же деле эффективная работа специалиста, оценивающего результаты лемматизации, возможна лишь при заполнении парадигм более чем на 80%.

3. Результаты исследования

По результатам эксперимента было выяснено, что большой процент слов, отсутствующих в словаре — это словоформы, написанные с ошибкой. Однако применение алгоритмов орфокоррекции для отсеивания ошибочных результатов здесь невозможно, так как многие новые слова отличаются на одну-две буквы от имеющихся в словаре. Не нарушая общности рассуждений предположим, что в словаре уже имеется слово «угол», но отсутствует слово «уголь». Тогда при анализе второго слова с учетом ошибок, слово «уголь» не будет добавлено к незнакомым словоформам, а увеличит встречаемость слово «угол» на единицу. Вторую большую группу составляли имена собственные. И лишь на последнем месте по количеству находятся слова из специальной или редко используемой общеупотребительной лексики. Для корпуса текстов из библиотеки Мошкова дополнительный хаос внесли тексты на белорусском и украинском языке. Точная оценка относительных размеров групп не проводилась в связи с большим объемом исследуемых корпусов.

Отдельную проблемы составляют слова, омонимичные как словарным, так и несловарным словоформам. В этом случае словоформа будет отнесена к словарным, а в парадигме соответствующей несловарной леммы будет «дырка». Данная проблема не исследовалась в нашей работе и нуждается в отдельной проработке.

В ходе исследования полученных результатов было выяснено, что практически ни одна группа гипотез, объединенных одним списком словоформ, не содержала в себе единственную и верную гипотезу. Среди прочего, это связано с тем, что в русском языке встречаются парадигмы, объединяющие один и тот же набор флексий, однако приписывающие

им различные наборы параметров. Так, для слова «админ», встретившегося на сайте www.bash.org.ru, порождались следующие леммы. «-» означает пустой постфикс. В скобках написаны словарные представители парадигмы.

Единственное число	**им. п.**	**род. п.**	**вин. п.**	**дат. п.**	**тв. п.**	**пред. п.**
АДМИН (ТЕЛЕФОН) м.р., неодуш	—	А	—	У	ОМ	Е
АДМИН (ТОН) м.р., неодуш	—	А	—	У	ОМ	Е
АДМИН м.р., неодуш	—	А	—	У	ОМ	Е/У
АДМИН (АКТИВИСТ) м.р., одуш	—	А	А	У	ОМ	Е
АДМИН (ОПЕР) м.р., одуш	—	А	А	У	ОМ	Е
Множественное число	**им. п.**	**род. п.**	**вин. п.**	**дат. п.**	**тв. п.**	**пред. п.**
АДМИН (ТЕЛЕФОН) м.р., неодуш	Ы	ОВ	Ы	АМ	АМИ	АХ
АДМИН (ТОН) м.р., неодуш	А/Ы	ОВ	А/Ы	АМ	АМИ	АХ
АДМИН м.р., неодуш	Ы	ОВ	Ы	АМ	АМИ	АХ
АДМИН (АКТИВИСТ) м.р., одуш	Ы	ОВ	ОВ	АМ	АМИ	АХ
АДМИН (ОПЕР) м.р., одуш	А/Ы	ОВ	ОВ	АМ	АМИ	АХ

Таким образом, даже один и тот же набор словоформ может быть различным образом размещен в различных парадигмах.

Для каждого из корпусов исследовалось количество лемм в зависимости от процента заполнения их парадигм. Дело в том, что по единственной словоформе довольно сложно корректно предсказать всю лемму. В связи с этим большое количество не полностью заполненных парадигм позволяет говорить о большом количестве ручной работы, которую придется проделать для лемматизации.

Для словарных слов была получена статистика, представленная на Рис. 1. На данном рисунке представлено распределение лемм по степени заполненности их парадигм для различных исследованных корпусов. Из нее видно, что с ростом объема корпуса растет и количество лемм, для которых встретилась большая часть их словоформ. То же самое можно видеть и на графике, представленном на Рис. 2. Здесь можно увидеть процент парадигм, заполненных не менее чем на 50%, в зависимости от логарифма объема корпуса. Данный график представляет собой, по всей видимости, сигмоиду, и из него видно, что на какой-то момент наступает насыщение. Для получения более точных результатов необходимо провести еще несколько экспериментов в промежуточных точках и точках на концах интервалов. Слова с не полностью заполненной парадигмой могут относиться, например, к специальной редковстречающейся лексике или к словам с дефектной парадигмой. В связи с этим дальнейшее увеличение объема корпуса не приведет к значительному изменению результатов. Так, на корпусе текстов из библиотеки Мошкова, процент парадигм, заполненных более чем на 50%, составил 99%.

Рис. 1. Зависимость относительной встречаемости парадигм от их заполнения (для словарных словоформ)

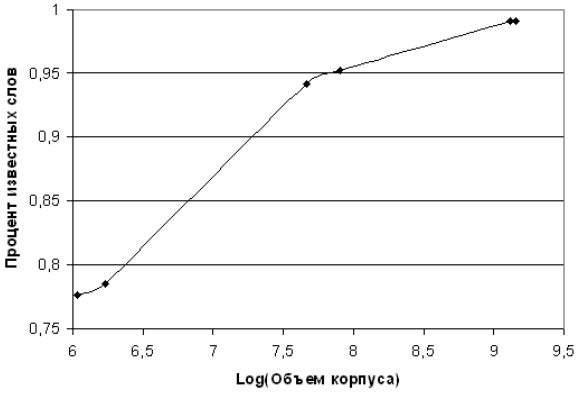

Рис. 2. Зависимость количества парадигм с заполнением >50% от логарифма объема базы (для словарных словоформ)

Рис. 3. показывает распределение лемм по степени заполненности их парадигм для несловарных словоформ. Легко видеть, что оно характеризуется противоположными тенденциями. Большая часть словоформ размещаются в парадигмах с заполнением менее 50%. При увеличении объема корпуса парадигмы постепенно перетекают из левой половины графика в правую, однако общая тенденция сохраняется. Это может быть объяснено, например, сохранением процента слов, содержащих ошибку, при увеличении объема базы. Также можно предположить, что с увеличением объема базы увеличива-

ется и число имен собственных, никогда до сих пор не встречавшихся, слов, относящихся к специальной лексике, сленгу, «новоязу» (образованию новых слов, значение которых ясно из контекста или используемой основы: «даунлоадить», «сторублируйте», «мазелин», …) и другим явлениям языка. Пик при значении 100% заполнения парадигмы связан с наличием неизменяемых слов, но в гораздо большей мере определяется словами с полностью заполненной парадигмой. Это говорит о том, что слова чаще всего либо встречаются лишь в нескольких формах на протяжении всего текста, либо (значительно реже) употребляются во всех своих словоформах. Следует заметить, что без предварительной фильтрации результатов, процент слабо заполненных парадигм был бы еще выше.

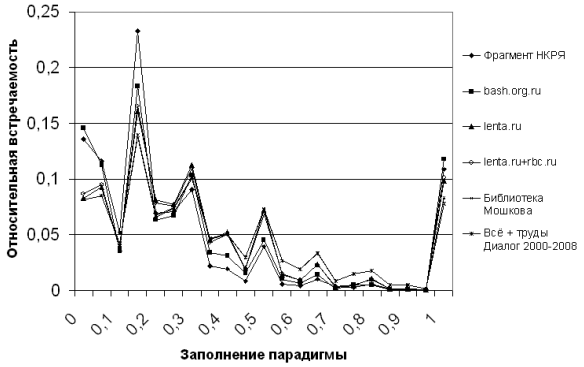

Рис. 3. Зависимость относительной встречаемости парадигм от их заполнения (для несловарных словоформ)

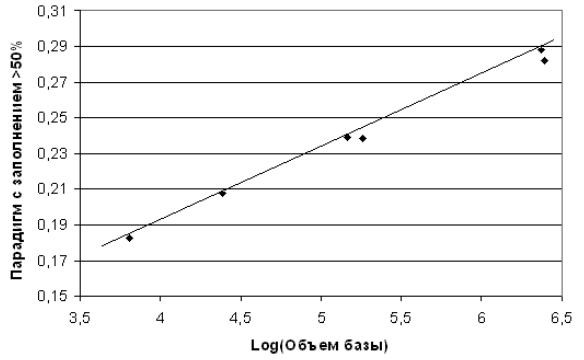

Рис. 4. Зависимость количества парадигм с заполнением >50% от логарифма объема базы (для несловарных словоформ)

Перетекание парадигм в область с заполнением >50% является довольно значительным (с 18 до 28%), но требует экспоненциального роста объема базы. Рис. 4. показывает зависимость количества таких парадигм от логарифма объема базы. Полученные данные неплохо апроксимируются прямой линией, из которой, однако, выпадают две точки, соответствующие относительно небольшому изменению объема базы. Можно предположить, что отклонение основывается на резком изменении используемой лексики. Следует заметить, что отсеивание белорусских и украинских текстов из библиотеки Мошкова могло изменить вид аппроксимирующей функции. Данный вопрос нуждается в дополнительных машинных экспериментах.

4. Обсуждение результатов

Анализ результатов показывает, что полностью автоматизировать процесс лемматизации несловарных словоформ на данный момент невозможно. Исключение может составить морфология, основанная на стемминге, когда нас больше интересует связь словоформы с нормальной формой или ее псевдооснова [4]. В этом случае можно пренебречь некоторыми нюансами. В противном случае, как это было показано выше, существует очень низкая вероятность получить по набору словоформ единственную включающую их парадигму.

Большое количество ошибок, встречающихся в любых текстах, зашумляет выход системы лемматизации и требует длительного ручного труда по отделению корректных вариантов от ошибочных. К счастью, возможностей для ошибки предоставляется очень много, и поэтому большинство ошибок встречаются один или два раза и отсеиваются на этапах фильтрации или кластеризации. Однако некоторые ошибочные словоформы могут войти в состав других парадигм, изменив тем самым результаты кластеризации не в лучшую сторону. Кроме того, у многих авторов существуют, так сказать, «любимые» ошибки, когда одна и та же ошибка допускается многократно в различных словоформах. Использование отредактированных источников должно облегчить труд, однако количество таких источников мало. Как показал эксперимент, даже в НКРЯ имеются отдельные лексические недостатки, не говоря об использовании разговорной лексики, не облегчающей труд системы лемматизации и специалистов.

И, наконец, дальнейшее развитие морфологического словаря упирается в экспоненциальный рост объема обучающего корпуса, что влечет за собой рост «шума» в результатах работы системы. Все вместе это ведет к постепенному замедлению работы по лемматизации. Общие трудозатраты на заполнение морфологического словаря имеют экспоненциальную форму.

Однако качественный скачок может быть получен путем применения ряда других подходов. Так, например, анализ окружения неизвестного слова существенно снижает омонимию, возникающую при выдвижении гипотез о принадлежности данного слова той или иной парадигме. Одним из методов, которые здесь можно применить, является метод триграмм [6]. Сам вопрос применения метода сня-

тия омонимии при работе с неизвестными словами в английском языке уже исследовался, например, в работе [7]).

Для получения лучших результатов можно предложить использовать специализированные корпуса научной направленности. Во-первых, количество ошибок в них существенно ниже, чем в большинстве современных литературных источников. А во-вторых, пополнение словарей ведется в основном за счет специальной лексики, которая как раз и расположена компактным образом в научных корпусах.

Резюмируя, следует сказать, что даже в текущем состоянии автоматизированный (а не автоматический) процесс лемматизации позволяет существенно сэкономить время специалиста, пополняющего морфологический словарь.

Литература

1. *Сидорова Е. А.* Многоцелевая словарная подсистема извлечения предметной лексики // Труды международного семинара Диалог'2008 «Компьютерная лингвистика и интеллектуальные технологии». М.: Наука, 2008. С. 475–481.
2. *Елкин С. В., Клышинский Э. С., Стекляннников С. Е.* Проблемы создания универсального морфосемантического словаря // Сб. трудов Международных конференций IEEE AIS'03 и CAD-2003, том 1, Дивноморское. 2003
3. *Сегалович И., Маслов М.* Русский морфологический анализ и синтез с генерацией моделей словоизменения для не описанных в словаре слов // Труды Международного семинара Диалог'98 по компьютерной лингвистике и ее приложениям. Казань: ООО «Хэтер», 1998. Т. 2. С. 547–552
4. *Ляшевская О. Н., Кобрицов Б. П., Сичинава Д. В.* Автоматизация построения словаря на материале массива несловарных словоформ // Интернет-математика 2007
5. *Черненьков Д. М.* Автоматизированное пополнение морфологического словаря на массиве текстовых документов // Труды научно-практического семинара «Новые информационные технологии-12». М.: МИЭМ, 2009. С. 138–141.
6. *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) // Интернет-математика-2005
7. *Mikheev A.* Automatic rule induction for unknown word guessing // Computational Linguistics, 23(3): 405–423, 1997