Обучение морфологического анализатора на большой электронной коллекции текстовых документов

Андреев А.М., Березкин Д.В., Симаков К.В.

НПЦ «ИНТЕЛТЕК ПЛЮС»
arka@inteltec.ru

Аннотация

В статье изложен метод и алгоритм обучения морфологического анализатора на основе большого текстового массива. В качестве учителя при обучении выступает морфологический анализатор словарного типа. Основная особенность обученного анализатора – способность выполнять разбор неизвестных слов. Проведен ряд экспериментов по оценке свойств алгоритма обучения, в частности свойства обобщения. Приведено сравнение работы обученного анализатора с анализатором словарного типа.

Введение

Основное назначение морфологического анализа текстов заключается в определении морфологических признаков слов текста и в определении канонических (начальных, нормализованных) форм слов. Задача является актуальной для любых компьютерных систем, обрабатывающих тексты на естественном языке (информационно-поисковые системы, электронные библиотеки, автоматические текстовые классификаторы, фильтры и др).

1. Основные методы морфологического анализа

Методы, применяемые в машинном морфологическом анализе, зависят от требований к точности выполняемого анализа, предъявляемых системой обработки. Так для поисковых систем [1], где основным требованием является выявление основы слова поискового запроса и определение его части речи, морфологический анализ сводится к определению вероятного словоизменительного окончания слова путем сопоставления с предопределенным словарем окончаний, где каждому окончанию с некоторой вероятностью приписана характерная часть речи. Зачастую анализируемому слову может быть приписано несколько вероятных окончаний из словаря окончаний. Также часто возникают ситуации, когда у слова вообще нет окончания (характерно для многих имен существительных, например, «язык», «анализ», «речь» и др.). Кроме того, для выделения неизменяемых групп слов (предлоги, союзы, местоимения, наречия и др.) необходимо иметь словарь таких исключений, где каждая группа представлена не окончанием, а полным перечнем слов. Ситуации, когда слово отсутствует в словаре исключений, либо ему приписывается нескольких окончаний, характерных разным частям речи, либо у слова вообще нет окончания, приводят к ошибкам морфологического анализа, которые сказываются на качестве работы всей системы в целом. Так если полнотекстовая поисковая система **ошибочно** выявила окончание «-ан» у слова «кран», то будет выполняться поиск по началу слова «кр*», в результате чего будут найдены документы, как содержащие слово «кран» или его возможные словоформы («краны», «краном» и т.д.), так и содержащие другие слова, начинающиеся на «кр*» («краб», «кров» и т.д.), что очевидно приведет к повышенному проценту нерелевантных документов, возвращаемых по результату поиска. Тем более такой подход не годится для более сложных задач обработки текста, таких, например, как выявление противоречий [12].

В противоположность описанному методу, базирующемуся на окончаниях, метод использующий словарь словообразовательных основ [2], [9], допускает ошибки значительно реже, поскольку опирается на предопределенный словарь основ, где каждой основе приписан набор постоянных морфологических признаков (часть речи, вид, переходность и др.). При анализе слова, выполняется поиск подходящей основы в словаре, по ней определяются постоянные морфологические признаки слова, а отсекая основу и суффиксы, выделяется окончание, по которому определяются переменные морфологические признаки (падеж, род, число и др.). Недостатком данного метода является зависимость качества разбора от полноты словаря основ, поскольку при отсутствии подходящей основы анализ слова либо прекращается безрезультатно, либо выполняется приближенным методом, основанным на окончаниях и рассмотренным ранее. Составление словаря основ выполняется вручную и представляет собой трудоемкую задачу. При этом словарь необходимо постоянно пополнять, поскольку в литературе часто появляются новые термины (например, в сфере высоких технологий). Так в работе [3] указывается на ежедневное пополнение словаря известного грамматического анализатора от Microsoft. К сожалению,

существующие в настоящий момент открытые словари страдают своей неполнотой, из-за некоммерческого характера этих разработок. Примером такого словаря может служить [4] (описывает ~90 тыс. слов), на основе которого построено множество морфологических анализаторов. В настоящее время проведенные эксперименты показали, что данный словарь позволяет выполнить разбор 48-50% слов, используемых в современных текстах, публикуемых на новостных сайтах Интернет.

Важной задачей, возлагаемой на процедуру морфологического анализа, является определение канонической (нормализованной) формы слова. Например, для причастий и деепричастий чаще всего канонической формой является инфинитивная форма глагола, от которого образовалось слово. Данная задача решается анализаторами на базе словаря словообразовательных основ путем введения ассоциативных таблиц суффиксов, где каждому суффиксу ставится в соответствие суффикс канонической формы. Для получения канонической формы необходимо выделить словообразовательную основу, суффикс и окончание слова, по ассоциативной таблице заменить суффикс суффиксом канонической формы и выполнить стыковку словообразовательной основы слова с найденным суффиксом канонической формы. Очевидно, для выполнения такой процедуры кроме таблицы суффиксов необходимо иметь мощный словарь словообразовательных основ (либо словарь канонических форм с явно выделенными суффиксами [4]). Составление и постоянное поддержание полноты такого словаря – сложная задача, требующая больших трудозатрат.

Еще одним методом морфологического анализа является метод аналогий, призванный устранить недостатки рассмотренных выше подходов: с одной стороны слова подлежат разбору, независимо от наличия/отсутствия у них окончания, с другой стороны слова подлежат разбору, независимо от наличия/отсутствия их основы в предопределенном словаре основ. Применение данного метода для автоматического машинного морфологического анализа текстов изложено в [5].

Идея метода заключается в том, что у подобных (аналогичных) по морфологическим признакам слов должны быть подобные концевые буквосочетания. Конец слова, в отличие от окончания, не выполняет словоизменительной функции, и в общем случае включает в себя окончание слова и некоторую концевую часть основы. Например, слова «крыльцом, лицом, винцом» обладают следующими морфологическим признаками: имя существительное, творительный падеж, средний род, единственное число. При этом все слова имеют общий конец «-цом», метод аналогии базируется на предположении, что любое слово с концовкой «-цом» с большой вероятностью будет обладать перечисленными морфологическими признаками.

В работе [5] кроме класса морфологических признаков каждому концу приписывалась длина окончания слова, так что кроме морфологических признаков появляется возможность определить словоизменительную основу, отсекая выявленное окончание. Для выделения канонических форм использовался мощный словарь словоизменительных основ слов и предопределенная таблица суффиксов, где каждому суффиксу соответствует суффикс канонической формы. Для выявления словоизменительной основы канонической формы слова, эта основа должна присутствовать в словаре. Для ее поиска в словоизменительной основе исходного слова выделялся гипотетический суффикс и заменялся по ассоциативной таблице на суффикс канонической формы. Если полученная после такой замены основа находилась в словаре, то делался вывод, о принадлежности анализируемого слова к найденной канонической форме, в противном случае перебор гипотетических суффиксов продолжался.

Недостатками предложенного в [5] подхода является следующие.

1. Для формирования словаря концов слов с приписанными им морфологическими классами и длинами окончаний необходимо вручную выполнить морфологический разбор огромного массива слов.
2. Для выделения канонической формы анализируемого слова необходимо иметь исчерпывающий словарь словоизменительных основ. Данный словарь может быть получен автоматически путем морфологического разбора большого массива слов (обучающая выборка) и отсечения у них окончаний. Тем не менее, итоговый словарь оказывается чрезмерно емким (~1 млн. основ). Наличие такого емкого словаря ограничивает возможности анализатора по производительности, поскольку процедура поиска словоизменительной основы канонической формы носит итеративный характер, связанный с перебором гипотетических суффиксов, где на каждой итерации выполняется обращение к емкому словарю основ.
3. Каноническая форма слова не может быть определена, если ее словоизменительная основа не присутствует в словаре основ. Теоретически проблемы в этом нет, поскольку словарь основ получается автоматически и при выявлении его неполноты процедура его формирования может быть запущена вновь. Но на практике для запуска автоматического формирования словаря должен быть изменен состав обучающей выборки, так чтобы в нее попала каноническая форма, выявление которой ранее было невозможно по причине отсутствия этой формы в обучающей выборке. По сути это означает, что человек-оператор должен вручную внести недостающую каноническую форму слова в обучающую выборку и запустить процедуру получения словаря основ. Такое вмешательство человека сводит на нет достоинства метода аналогии, и существенно затрудняет практическое использование морфологического анализатора в рамках реальной системы.

2. Развитие метода аналогии

Учитывая недостатки описанных существующих подходов морфологического анализа, был разработан метод, в основу которого положен метод аналогии, имеющий следующие отличия:

- обучение анализатора выполняется автоматически,
- для определения канонической формы слова автоматически синтезируется словарь концевых буквосочетаний канонических форм, который существенно меньше исчерпывающего словаря словоизменительных основ,
- при синтезе канонической формы у слова отсекается характерное для его морфологических признаков концевое буквосочетание и присоединяется концевое буквосочетание канонической формы, при этом не приходится выполнять итеративные обращения к емкому словарю основ и перебирать вероятные основы анализируемого слова, что положительно сказывается на производительности анализатора.

2.1 Обучение анализатора, используя учителя

За рубежом распространена практика использования размеченных корпусов текстов для обучения текстовых анализаторов различного вида: морфологических [6], синтаксических [7] и семантических [8]. Разметка текста выполняется вручную человеком, где те или иные элементы текста снабжаются специальными метками, определяющими их свойства. В России также имеются размеченные корпусы (например, [10]), но они, к сожалению, открыты для конечного пользователя, а не для исследователя или разработчика. Так [10] предлагает интерактивное взаимодействие с корпусом посредством Web, но не предоставляет прямой доступ к хранимым текстам.

Метод аналогии в [5] в какой-то мере повторяет этот подход, используя разобранный человеком массив слов. Идея предложенного в настоящей работе подхода заключается в том, чтобы использовать для этих целей не человека, а другой морфологический анализатор, выступающий в роли учителя.

Для предъявления требований к учителю дадим строгое описание работы обученного морфологического анализатора и процедуре его обучения. Пусть имеется множество образцов P (от англ. patterns), представляющих собой характерные концевые буквосочетания слов. Так для слов «крыльцом, лицом, винцом» образцом является *«-цом»*=$p_i \in P$. Пусть также имеется множество морфологических шаблонов T (от англ. templates), представляющих собой группы морфологических признаков, которые могут быть приписаны классу слов с общим образцом, для приведенного примера шаблоном является «имя существительное, творительный падеж, средний род, единственное число» или сокращенно *«сущ\|Т\|С\|ЕД\|»*=$t_j \in T$. Одному и тому же образцу может соответствовать несколько шаблонов и наоборот. Морфологический анализатор, действующий по методу аналогии, может быть описан алгебраической моделью $M =< P, T, R >$, где R – отношение инцидентности на $P \times T$, всякая пара $(p,t) \in R$ описывает класс слов с концом p и морфологическими признаками t. Пусть W - множество всех возможных словоформ всех слов естественного языка. Обученный морфологический анализатор, действующий по методу аналогии, для любого слова $w \in W$ формирует подмножество $R_w \subset R$, где для любой пары $(p_i,t_j) \in R_w$ образец p_i является концом слова w, а шаблон t_j является возможным набором его морфологических признаков. Анализатор-учитель в отличие от обученного морфологического анализатора должен распознавать некоторое подмножество слов $W_t \subset W$, в котором для каждой допустимой пары $(p,t) \in R$ существует, по крайней мере, два слова. При удовлетворении этому требованию, учитель будет в состоянии сформировать M на основе обучающей выборки W_t, выявляя для каждого слова w_i с некоторым морфологическим шаблоном t парное слово w_j с тем же шаблоном t, так чтобы пара (w_i,w_j) имела максимальную длину общего концевого буквосочетания, которое может быть объявлено образцом p.

2.2 Определение канонической формы

Предположение, на котором основан метод определения канонической формы, формулируется следующим образом: у словоформ w_i и w_j с одинаковым концевым буквосочетанием p_w и морфологическим шаблоном t канонические формы c_i и c_j также имеют некоторое одинаковое концевое буквосочетание p_c. Для приводимого примера слов *«крыльцом, лицом, винцом»* каноническими формами являются *«крыльцо, лицо, винцо»*, для которых образцом является p_c=*«-цо»*.

Для отражения этого факта в модель M необходимо добавить еще одно отношение R_c на $R \times P$, так что каждая пара $((p,t), p_c) \in R_c$ отражает тот факт, что у любого слова, соответствующего образцу p и имеющему морфологический шаблон t, каноническая форма соответствует образцу p_c. Поэтому морфологический анализатор, действующий по методу аналогии и способный определять канонические формы, описывается следующим образом $M =< P, T, R, R_c >$. Получение канонической формы таким анализатором для некоторого слова w сводится к следующим шагам:

- определение образца p и морфологического шаблона t, которым соответствует слово w,
- определение образца p_c канонической формы, так что $((p,t), p_c) \in R_c$,
- замена образца p в слове w на образец p_c с образованием канонической формы c.

Строго говоря, одному p может соответствовать несколько p_c образцов канонических форм. Например, для слов: *«топчась, корчась, мелочась, прячась»*, соответствующих образцу p=*«-чась»*, канонические формы будут *«топтаться, корчиться, мелочиться, прятаться»*, с образцами p_1=*«-таться»* и p_2= *«-читься»*. В этом случае при анализе, например, слова *«топчась»* анализатор сформи-

рует две канонические формы c_1=*«топтаться»* и c_2=*«топчиться»*, вторая из которых является **ошибочной**. Для минимизации количества таких случав в процедуру обучения необходимо ввести дополнительное ограничение. Так что при формировании образца p на основе некоторой пары слов (w_i,w_j) необходимо брать **не** максимальную длину общего для этих слов концевого буквосочетания. Длина образца p должна быть такой, чтобы, отсекая его от обоих слов, получались такие начальные буквосочетания слов $b_i=w_i - p$ и $b_j=w_j - p$, чтобы при отсечении их от начала канонических форм c_i и c_j получался один и тот же образец канонической формы $p_c=c_i - b_i=c_j - b_j$ (здесь знаком «-» минус обозначена операция отсечения символов с конца слова). В проводимых экспериментах введение такого ограничения в процедуру обучения анализатора приводило к сокращению процента неоднозначно синтезируемых канонических форм с 15% до 0,1-0,3%, т.е. итоговые анализаторы в 99,7 - 99,9% случаев синтезировали однозначные канонические формы.

При выполнении анализа массива текстов (а не отдельно взятого слова), возникает дополнительная возможность отфильтровать ложные канонические формы статистическими методами. Так, если для слова w_1 формируется множество гипотетических канонических форм C_1, а для слова w_2 – множество C_2, то, если w_1 и w_2 имеют общую каноническую форму c, пересечение $C=C_1 \cap C_2$ будет либо содержать только c, либо некоторое число общих для C_1 и C_2 элементов, среди которых присутствует c. Поэтому можно выполнить замену множеств C_1 и C_2 на C для обоих слов. Итеративно выполняя подобную замену для всех слов анализируемого массива, в которых $C_1 \cap C_2 \neq \varnothing$, множества C_1 и C_2 будут сокращаться, в идеале стремясь к $C_{lim}=\{c\}$. В проведенных экспериментах описанная процедура фильтрации сократила число неоднозначно генерируемых канонических форм на 20-45% (в зависимости от анализируемого массива) от их изначального числа. Множества типа C_1 и C_2, выдаваемые обученным анализатором, могут использоваться непосредственно без однозначного выявления в них канонической формы $c \in C_i$ в задачах поиска, классификации [11], реферирования, сопоставления синтаксических шаблонов [1]. Как правило, в таких случаях достаточно иметь возможность относить разные словоформы одного слова к одной и той же канонической форме. Так для выявления того факта, что словоформы w_1 и w_2 имеют общую каноническую форму достаточно удостовериться, что $C_1 \cap C_2 \neq \varnothing$. В проведенных экспериментах данное предположение не привело ни к одной ошибке.

2.3 Алгоритм обучения

Учитывая описанные особенности, процедура обучения сводится к следующим шагам, отраженным на рис. 1. В качестве учителя выступает морфологический анализатор словарного типа. На вход анализатору подаются слова из текстов обучающей электронной коллекции, при этом тексты не наделены какой-либо разметкой. Поскольку словарь учителя может быть неполным и слова из текстов берутся «как есть» (т.е. могут иметь опечатки и ошибки) некоторые из них могут быть не разобраны учителем, в таком случае они записываются в журнал (Log1 на рис. 1) для проведения дальнейшего тестирования обученного анализатора. Успешно разобранные учителем слова записываются в другой журнал (Log2 на рис. 1) и группируются по морфологическим шаблонам.

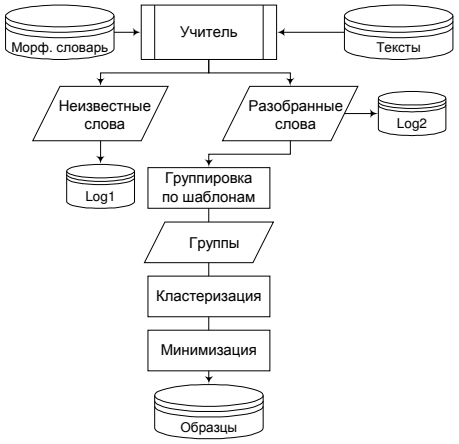


Рис. 1. Функциональная схема обучения.

На рис. 1 приведены примеры таких групп для шаблонов t_i=*«сущ|Т|С|ЕД|»* и t_j=*«прг|М|ЕД|Д|»*. Далее эти группы подвергаются кластеризации, так чтобы внутри них выявить кластеры слов, отвечающих общим образцам, т.е. имеющие общие концевые буквосочетания, что по сути означает формирование образцов. Для этого внутри каждой группы перебираются все возможные пары (w_i,w_j) слов на предмет возможной генерации образца на их основе. Проверка основана на четырех критериях, все кроме первого являются принципиальными отличиями данной работы, от подхода, изложенного в [5].

1. Первым критерием генерации образца является максимальная длина общего для пары (w_i,w_j) слов концевого буквосочетания. Т.е., беря в группе некоторое слово w_i, в этой же группе выполняется поиск парного ему слова w_j, такого чтобы их общее концевое буквосочетание было максимальной длины.
2. Если пара (w_i,w_j) удовлетворяет первому критерию, то необходимо проверить совпадают ли при этом образцы канонических форм этих слов. Совпадение образцов канонических форм является вторым критерием.

В качестве примера рассмотрим слово w_i=*«колеблется»*. По первому критерию для формирования максимально длинного образца подходит парное слово w_j=*«зыблется»*, в этом случае в качестве образца можно принять фрагмент p=*«-блется»*, как наиболее длинное концевое буквосочетание, общее для обоих слов. Канонические формы данных слов соответственно c_i=*«колебаться»* и c_j=*«зыбиться»*. Для получения образцов канониче-

ских форм обоих слов необходимо отсечь от них p, в результате останутся начальные буквосочетания b_i=*«коле-»* и b_j=*«зы-»*. Далее отсекая эти сочетания от начал обоих канонических форм, получаем образцы p_{ci}=*«-баться»* и p_{cj}=*«-биться»*. Как видно, образцы канонических форм не совпадают, поэтому исходное предположение о том, что для генерации образца на основе слова w_i=*«колеблется»* подходит парное слово w_j=*«зыблется»*, является **неверным** и необходимо искать другое слово, с меньшей длиной образца. Таким словом для данного примера является w_j=*«треплется»*, при этом генерируется образец p=*«-лется»*. Каноническая форма для этого слова – c_j=*«трепаться»*, отсекая от нее начальное буквосочетание *«треп-»* получаем образец p_{cj}=*«-аться»*, что совпадает *с* p_{ci}=*«-аться»*, полученным при отсечении от c_j=*«колебаться»* начального буквосочетания *«колеб-»*.

Для сокращения процента ошибок в работе обученного анализатора на этапе кластеризации пришлось также добавить два дополнительных критерия, учитывающие следующую особенность его работы. Обученный анализатор располагает образцами $P=\{p_i\}$, при анализе слова w выполняется поиск в P наиболее длинного p_i, совпадающего с концом слова w, при этом слово может удовлетворять и другим образцам $P_w \subset P$ меньшей длины, чем p, но они в расчет не берутся. Поэтому на этапе кластеризации необходимо обеспечить такую генерацию образцов, чтобы генерируемый образец p для слова w в последствии не попадал в P_w при анализе этого же слова уже обученным анализатором. В противном случае при анализе для слова всегда будет находиться другой ошибочный образец, а правильный образец будет по ошибке отбрасываться в P_w. Учитывая эту особенность вводятся следующие коррективы.

3. Прежде чем генерировать образец p на основе слова w_i и некоторого парного ему слова необходимо проверить, не удовлетворяет ли слово w_i уже существующим образцам в других группах. Если для данного слова удается найти уже существующие образцы P_e, то из их необходимо выбрать образец p_e максимальной длины и потребовать, чтобы генерируемый на данном шаге образец p был длиннее p_e по крайней мере на один символ. Если не вводить указанной коррективы, при анализе слова w_i будет ошибочно обнаружен образец p_e (потому что он длиннее, чем p), который изначально относился к другой группе слов, т.е. к другому морфологическому шаблону, что приведет к ошибке анализа.
4. После генерации образца p на основе слова w_i и некоторого парного ему слова необходимо проверить, не существует ли среди уже проанализированных слов, относящихся к другим морфологическим шаблонам, такого слова w_e, которое также удовлетворяло бы образцу p. Если такое слово найдено, то необходимо убедиться, что его образец p_e длиннее, чем p, в противном случае необходимо отказаться от принятия p и продолжить попытки генерации более короткого образца на основе исходного слова w_i. Если не предпринимать указанных действий, то при работе обученного анализатора разбор слова w_e приведет к тому, что оно будет отнесено к ошибочно сгенерированному образцу p, а не к образцу p_e, полученному на основе w_e в процессе обучения.

После кластеризации групп слов и генерации образцов выполняется этап минимизации, заключающийся в допустимом сокращении длин образцов. Такая процедура уместна, поскольку основным критерием генерации образцов является максимизация их длин. В результате чего образуется избыточное множество образцов, в котором встречаются образцы отличающиеся в первых 1-2 символах и отвечающих одним и тем же морфологическим шаблонам. Задача этапа минимизации – выявить такие повторения и объединить их в общие кластеры с минимальной длиной образца. Кроме того, возможны ситуации, когда кандидатов на объединение вообще нет, в таких случаях образец может быть просто укорочен. Проведенные эксперименты показали, что на этапе минимизации объем словаря образцов P, полученный в результате кластеризации, может быть сокращен примерно в 2 раза.

3. Результаты экспериментов

Обучение морфологического анализатора проводилось на электронной коллекции текстовых сообщений средств массовой информации «Обзор СМИ», функционирующей в настоящий момент в Совете Федерации Российской Федерации под управлением ОСУБД Odb-Jupiter 4.0. Коллекция ведется с 2002 года и к настоящему моменту насчитывает ~400 тыс. новостных сообщений общеполитического содержания. Словарный объем коллекции составляет ~1 млн. словоформ русского языка.

В качестве учителя использовался морфологический анализатор словарного типа, в основе которого лежит грамматический словарь [4]. Данный словарь включает ~90 тыс. канонических форм слов, эксперименты показали, что при анализе тестовой коллекции «Обзор СМИ» было задействовано ~80 тыс., что позволяет судить о степени репрезентативности данного текстового массива. Анализатором-учителем удалось разобрать только 48-50% словоформ исходной тестовой коллекции (~500 тыс. словоформ), остальные отброшенные на этапе обучения слова оказались не учтенными словарем Зализняка [4] и являлись: новыми терминами, именами собственными, сокращениями и обозначениями, словами с орфографическими ошибками. Все они записывались в журнал (Log1 на рис. 1) для дальнейшей проверки полноты обученного анализатора.

3.1 Свойства обученного анализатора

Для оценки зависимости свойств анализатора от емкости обучающей выборки было проведено четыре эксперимента, в которых принимало участие 50%, 66%, 75% и 100% словоформ от всего обучающего массива.

Табл. 1. Распределение слов и шаблонов по образцам.

	50%			66%			75%			100%		
L_p	$P_{1/2}$	$W_{1/2}$	$T_{1/2}$	$P_{2/3}$	$W_{2/3}$	$T_{2/3}$	$P_{3/4}$	$W_{3/4}$	$T_{3/4}$	P_1	W_1	T_1
0	-	23099	-	-	29369	-	-	32517	-	-	35509	-
1	30	120	1.36	30	116	1.66	30	152	1.83	30	121	1.80
2	335	1678	1.63	367	1574	1.71	362	2008	1.77	387	2011	1.95
3	3834	32564	1.43	4065	42743	1.53	4204	45805	1.58	4378	69239	1.72
4	9085	37849	1.34	10069	52230	1.44	10468	61535	1.48	11150	96734	1.58
5	16750	63659	1.31	19235	80630	1.38	20091	98507	1.40	22167	142994	1.47
6	18317	61457	1.39	21751	88837	1.47	22860	97528	1.48	25110	138803	1.54
7	17973	52709	1.47	20552	71112	1.57	21512	79540	1.60	21946	108476	1.65
8	14606	40148	1.55	16531	54363	1.64	16868	58620	1.69	15796	70555	1.76
9	11679	27523	1.57	13090	35663	1.67	13720	38900	1.72	12356	42642	1.84
10	7405	16431	1.58	8451	21483	1.70	8657	23056	1.75	7343	22809	1.93
11	4754	10124	1.57	5235	12421	1.69	5523	13989	1.77	4335	12798	1.99
12	2550	5166	1.58	2975	6624	1.68	2961	7148	1.76	2322	6258	1.98
13	1408	2664	1.54	1521	3217	1.66	1606	3595	1.75	1154	3003	2.02
14	740	1273	1.48	812	1572	1.61	904	1844	1.67	612	1436	1.93
15	349	559	1.41	399	679	1.53	430	764	1.57	288	609	1.78
16	153	215	1.31	187	332	1.59	208	339	1.58	168	328	1.90
17	89	137	1.52	110	187	1.64	124	230	1.77	98	200	2.01
18	46	75	1.54	54	109	1.77	46	73	1.56	42	98	2.07
19	17	27	1.58	22	34	1.45	24	45	1.75	9	21	2.33
20	8	12	1.50	2	4	2.00	7	13	1.85	3	6	2.00
∑	110128	-	-	125458	-	-	130605	-	-	129694	-	-

В таблице 1 приведено распределение словоформ и шаблонов по образцам разной длины. В первой колонке (L_p) приведены длины полученных образцов. Колонки $P_{1/2}$, $P_{2/3}$, $P_{3/4}$ и P_1 содержат количества сгенерированных образцов для каждой длины (индекс *1/2* соответствует эксперименту с участием 50% словоформ от всей обучающей выборки, *2/3* – 66%, *3/4* – 75%, *1* – 100%). Колонки $W_{1/2}$, $W_{2/3}$, $W_{3/4}$ и W_1 содержат количества словоформ, участвующих в генерации образцов каждой длины. Колонки $T_{1/2}$, $T_{2/3}$, $T_{3/4}$ и T_1 содержат среднее количество морфологических шаблонов, приходящихся на один образец каждой длины.

Первой строке (длина образца 0 букв) соответствуют словоформы, у которых обучающий алгоритм не смог выделить общих концевых буквосочетаний в рамках общих морфологических шаблонов. К таким словам в большинстве своем относятся наречия, местоимения, предлог и союзы, представляющие собой потенциальный словарь исключений. Как видно из табл. 1 предельное количество таких слов составило 35509 в последнем эксперименте.

В результате обучения на полной коллекции (см. колонки для 100%) было сгенерировано 129696 образца. Максимальная длина образца - 20 букв. В силу морфологической неоднозначности многие слова относились к нескольким морфологическим шаблонам, для каждого такого слова учитывались все его повторы. В результате в процедуре обучения принимало участие ~750 тыс. словоформ (суммируются значения столбца W_1).

Из таблицы видно, что наибольшее число образцов имеет длину от 3 до 9 букв, этим образцам соответствует 89% всех словоформ. Этим же образцам соответствует минимальная морфологическая неоднозначность (столбцы $T_{1/2}$, $T_{2/3}$, $T_{3/4}$ и T_1), так для образцов длиной 5 букв среднее число морфологических шаблонов, приписываемых слову в результате анализа равно T_1=1,47. Данное наблюдение отвечает реальной действительности, так, например, у многих существительных совпадают формы именительного и винительного падежа и/или формы родительного и винительного, то же самое относится к именам прилагательным.

Важно ответить на вопрос, какова зависимость объема словаря образцов от объема обучающей выборки, т.е. необходимо выявить, сходится ли к некоторому пределу результат обучения. Если объем словаря будет расти линейно вместе с ростом количества словоформ обучающей выборки, то с одной стороны анализатор будет лишен практической ценности, а с другой стороны это будет означать отсутствие способности обобщения у обучающего алгоритма. Объем словаря V_i для i-ого эксперимента оценивается общей суммой полученных образцов (строка $\sum$ табл. 1), к которой прибавляется количество слов-исключений (первая строка табл. 1). Так для эксперимента с половиной обучающей выборки (колонки с заголовком 50% в табл. 1) $V_{1/2}$ = 110128 + 23099=133227, аналогично получаются: $V_{2/3}$=154827, $V_{3/4}$=163122 и V_1=165203. На рис. 2 приведен график, отражающий зависимость объема словаря обученного анализатора от объема обучающей выборки. На графике объем обучающей выборки выражен в относительных величинах, для получения абсолютных значений, т.е. числа словоформ, участвующих при обучении в i-ом эксперименте, можно просуммировать соответствующую колонку W_i таблицы 1. Из рис. 2 видно, что объем итогового словаря растет с увеличением объема обучающей выборки, но эта зависимость имеет нелинейный характер, более того можно утверждать, что объем результата обучения сходится к некоторому пределу, что позволяет судить о способности разработанного обучающего алгоритма к обобщению, а также о наличии некоторого предела, начиная с которого обучение прекратится, поскольку полученный словарь будет охватывать все словоформы естественного языка.

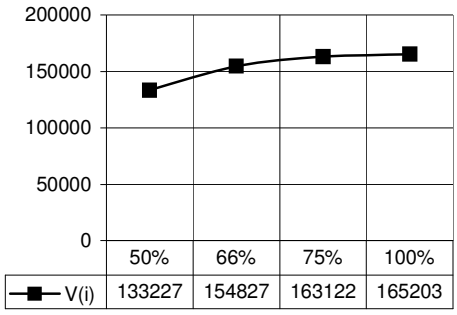


Рис. 2. Зависимость объема словаря обученного анализатора от объема обучающей выборки.

Еще одним интересным наблюдением, позволяющем судить о способности алгоритма к обобщению, служит перераспределение обучающих словоформ по образцам с длиной от 8 до 15 – при увеличении общего числа обучающих словоформ. Они стремятся распределяться по образцам с меньшей длиной. Эта тенденция отражена на рис. 3 для эксперимента с 75% и 100% объемами обучающей выборки.

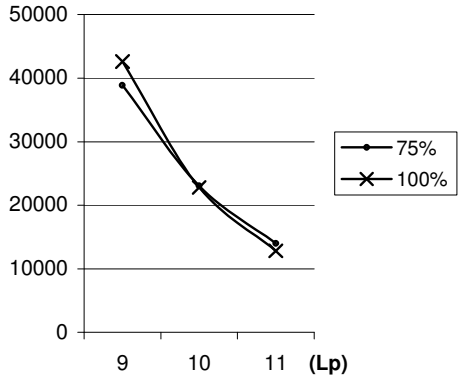


Рис. 3. Перераспределение обучающих словоформ по длинам образцов.

Как видно из рис. 3, чем больше объем обучающей выборки, тем меньшее число словоформ участвует в генерации образцов с длиной более 10. Т.е. площадь под графиком 100% меньше площади под графиком 75% на диапазоне от 10 до 11, и наоборот площадь под графиком 100% больше площади под графиком 75% на диапазоне от 9 до 10. Это позволяет сделать вывод, что при увеличении количества обучающих словоформ алгоритм начинает генерировать для них больше общих образцов меньшей длины, что вы итоге положительно сказывается на объеме итогового словаря образцов, т.к. в него попадает больше коротких образцов и меньше длинных.

3.2 Оценка точности и полноты обученного анализатора

Кроме оценки способности к обобщению обучающего алгоритма необходимо оценить точность и полноту обученного анализатора. Для оценки точности использовалась следующая схема (см. рис. 4).

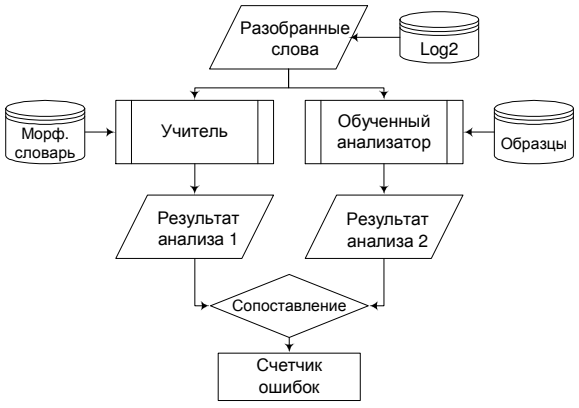
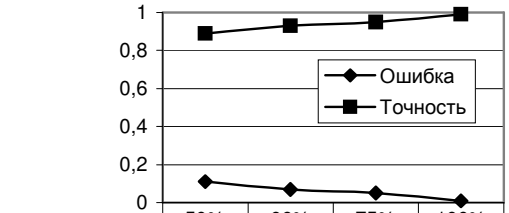


Рис. 4. Схема оценки точности анализатора.

Слова из журнала Log2, в который при обучении сохранялись удачно разобранные учителем словоформы (см. рис. 1), поступают на вход обученного анализатора и учителя. Оба результата разбора сопоставляются и в случае выявления расхождения –

фиксируются в счетчике ошибок. Данному эксперименту подвергались все четыре обученных анализатора, полученные соответственно при использовании 50%, 66%, 75% и 100% словоформ от всего обучающего массива. Во всех четырех случаях использовался полный список разобранных словоформ из последнего эксперимента с 100% обучающей выборкой, что позволило дополнительно проверить первые три анализатора на способность разбирать слова, не участвовавшие при их обучении. Относительная ошибка анализатора вычислялась делением содержимого счетчика ошибок на общее количество словоформ из Log2, которое, как упоминалось ранее, составляет ~ 500 тыс. словоформ. Результаты оценки точности и ошибки приведены на рис. 5.

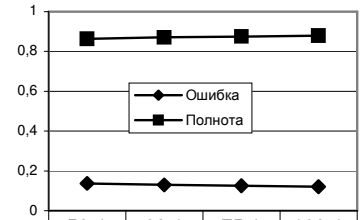


	50%	66%	75%	100%
Ошибка	0,11	0,07	0,05	0,01
Точность	0,89	0,93	0,95	0,99

Рис. 5. Оценка точности обученного анализатора.

Точность анализатора определяется как дополнение относительной ошибки до 1. Из рис. 5 видно, что с ростом объема обучающей выборки, процент ошибочно разобранных словоформ стремится к 0, и наоборот точность разбора в пределе стремится к 1. Рис. 5 также позволяет сравнить обученный анализатор с анализатором словарного типа, в роли которого выступает учитель. Анализатор, обученный на 50% исходной выборки, дает правильный ответ в 89% случаев, при этом следует учитывать, что половина тестируемых слов для него вообще не известна, поскольку эти слова не присутствовали при его обучении. Анализатор, обученный на 75% исходной выборки, дает ошибочный ответ лишь на 5% анализируемых слов, при том, что 25% анализируемых слов ему не известны. В таких же условиях анализатор словарного типа выдал бы 25% ошибок, поскольку он не в состоянии анализировать слова, о которых ему ничего не известно. В этом отношении анализатор, обученный по разработанному алгоритму, имеет существенное преимущество – возможность разбора неизвестных слов. Более точно эту возможность позволяет отразить такой параметр, как полнота разбора.

В данной работе для оценки полноты разбора использовались словоформы, изначально отбракованные учителем в Log1 (см. рис. 1). Иными словами проверялась способность обученных анализаторов разбирать словоформы, которые не смог разобрать учитель, т.е. абсолютно неизвестные. Число таких словоформ составило ~500 тыс., все они подавались на вход каждому из четырех обученных анализаторов, для каждого анализатора фиксировалось количество словоформ, которые он не в состоянии был разобрать. Результаты этого эксперимента представлены на рис. 6.



	50%	66%	75%	100%
Ошибка	0,137	0,13	0,126	0,121
Полнота	0,863	0,87	0,874	0,879

Рис. 6. Оценка полноты обученного анализатора.

На рис. 6 приведена относительная ошибка, полученная делением числа неразобранных словоформ на общее число словоформ теста. Полнота в данном случае определяется как дополнение ошибки до 1. Как видно из рис. 6, тенденция к росту полноты с увеличением объема обучающей выборки просматривается, хотя и не так явно, как в случае с точностью. Такое поведение обученных анализаторов объясняется отсутствием эталона, с которым можно было бы сопоставлять выдаваемые ими результаты. Единственным критерием оценки качества разбора в данном эксперименте является фиксирование факта способности/неспособности разбора анализаторами неизвестных слов. Если в качестве эталона использовать анализатор, обученный на 100% обучающей выборки, то возможно применять схему тестирования, представленную на рис. 4, но такая схема не может быть гарантированно объективной, поэтому в данной работе эти результаты не представлены.

Заключение

В работе предложена и проверена экспериментально модификация метода аналогии морфологического анализа слов. Разработан алгоритм автоматического обучения морфологического анализатора с применением учителя. Экспериментально показана сходимость процесса обучения при росте объема обучающей выборки. Сходимость процесса обучения вместе с высокими показателями точности обученного анализатора позволяют сделать вывод о практической применимости предложенного подхода.

В работе достигнута основная цель – разработан алгоритм, позволяющий получить морфологический анализатор, не уступающий по точности разбора анализаторам словарного типа, но существенно превосходящий их в отношении полноты анализа. Это достоинство отражается в способности обученного

анализатора разбирать новые ранее неизвестные слова.

Предложенный метод синтеза канонической формы в 99,7% случаев дает однозначный и правильный результат, а в 100% случаев результат содержит правильно синтезированную форму. Этого достаточно для применения данного подхода как в задачах идентификации слов при анализе групп текстов (реферирование, поиск и др.), так и в более требовательных к правильности написания слов задачах, например, при автоматической генерации названий элементов пользовательского интерфейса.

Литература

[1] Брик А.В. Исследование и разработка вероятностных методов синтаксического анализа текста на естественном языке. Диссертация на соискание ученой степени к.т.н. - М.: МГТУ им. Н.Э. Баумана, 2002.

[2] http://starling.rinet.ru/morph.htm

[3] Stephen D. Richardson, William B. Dolan, Lucy Vanderwende, MindNet: acquiring and structuring semantic information from text. Proceeding of ACL - Colling 1998.

[4] Зализняк А. А. Грамматический словарь русского языка (словоизменение). 2-е изд., М., "Русский язык", 1980.

[5] Г.Г. Белоногов, Ю.П. Калинин, А.А. Хорошилов, Компьютерная лингвистика и перспективные информационные технологии - М.: Русский мир, 2004.

[6] Jan Hajič, Morphological tagging: Data vs. Dictionaries, Proceedings of the first conference on North American chapter of the Association for Computational Linguistics, 2000.

[7] Shlomo Argamon, Ido Dagan, Yuval Krymolowski, A memory-based approach to learning shallow natural language patterns. ACM'98.

[8] Ted Pedersen, A Simple Approach to Building Ensembles of Naive Bayesian Classifiers for Word Sense Disambiguation. ACM'2000.

[9] И. Сегалович, М. Маслов, Русский морфологический анализ и синтез с генерацией моделей словоизменения для не описанных в словаре слов, http://company.yandex.ru/articles/article1.html

[10] Корпус русского языка http://www.ruscorpora.ru

[11] Андреев А.М., Березкин Д.В., Морозов В.В., Симаков К. В. Автоматическая классификация текстовых документов с использованием нейросетевых алгоритмов и семантического анализа. 5-ая Всероссийская научная конференция RCDL'2003.

[12] Андреев А.М., Березкин Д.В., Симаков К. В. Особенности проектирования модели и онтологии предметной области для поиска противоречий в правовых электронных библиотеках. 6-ая Всероссийская научная конференция RCDL'2004.

Unsupervised learning of morphological analyser using huge corpus of natural texts

This paper presents the method and algorithm for unsupervised learning of morphological analyser meaning that no human interactions are needed to control learning process. The algorithm uses only two source of knowledge about natural language. The first one is a huge collection of unrestricted natural texts. The second one is a teacher that is a dictionary-based morphological analyser. The main feature of trained analyser is that it can analyse some words that teacher can't.

We have carried out several experiments to make estimation of our learning algorithm and to evaluate properties of trained analyser. In particular we estimate the ability of learning algorithm to generalize words. This feature allows trained analyser to process unknown words.

Also we have assessed precision of trained analyser comparing results of its work with results of dictionary-based analyser.