УЧЕНЫЕ ЗАПИСКИ КАЗАНСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА
Том 151, кн. 3 Физико-математические науки *2009*

УДК 004.81

ЭВРИСТИКИ ДЛЯ УЛУЧШЕНИЯ РАБОТЫ ЧАСТИЧНОГО СИНТАКСИЧЕСКОГО АНАЛИЗАТОРА

В.А. Буштедт, В.Н. Поляков

Аннотация

В статье описана работа частичного синтаксического анализатора с использованием эвристик, которые позволяют сократить количество ложно выявленных при первоначальном анализе синтаксических конструкций (чанков). В предыдущих работах авторов было выявлено, что в русском языке на обнаружение чанков очень большое воздействие оказывают явления омонимии и полисемии. Под ложно выявленными чанками понимаются такие, которые были обнаружены частичным синтаксическим анализатором, но не являются на самом деле верными. Метод поиска чанков с применением этих эвристик получил название «Right-Chunk 4». Приведена формальная постановка задачи. Выполнена компьютерная реализация метода поиска чанков, результатом которой является программный комплекс «Chunk-Creator 4». Проведена оценка эффективности предложенного метода поиска чанков.

Ключевые слова: искусственный интеллект, компьютерная лингвистика, парсинг, синтаксический анализ, чанкинг.

Введение

Задача синтаксического анализа является одной из сложных задач компьютерной лингвистики и искусственного интеллекта. Исследование данной проблемы началось еще в 60-е годы XX века. Были созданы различные системы, которые позволяли проводить синтаксический анализ предложений на естественном языке, но сильного развития эти системы не получили, так как имели не очень высокие показатели точности анализа [1, 2].

Позднее задачу синтаксического анализа начали решать в составе задачи автоматического анализа текста на естественном языке. Но стоит отметить, что до недавнего времени большое число исследователей сходились во мнении о нецелесообразности введения модуля синтаксического разбора в системы автоматического анализа текста ввиду сложности реализации этой идеи [3]. Однако оказалось, что, несмотря на ограниченную точность синтаксических анализаторов, их использование способно заметно повысить качество таких систем в случае комбинирования с известными статистическими методами [4, 5].

К решению этой проблемы существуют три подхода: формально-грамматический, эвристический и вероятностно-статистический [6].

Первый направлен на создание сложных систем правил, которые позволяли бы в каждом конкретном случае принимать решение в пользу той или иной синтаксической структуры; второй – на выявление таких эвристик, которые бы позволяли получать наиболее приближенные к реальности результаты работы синтаксического анализатора; третий — на сбор статистики встречаемости различных структур в похожем контексте, на основе которой и принимается решение о выборе варианта структуры.

В настоящей работе рассматривается формально-грамматический и эвристический подходы.

Вследствие того, что достичь удовлетворительных результатов при полном синтаксическом анализе пока не представляется возможным, нами было принято решение о проведении исследований, связанных с работой частичного синтаксического анализа, так называемого «Чанкера» (от англ. слова «chunk» – глыба, ломоть, то есть нечто грубое и общее, в смысле частичного синтаксического анализа по сравнению с полным).

При синтаксическом анализе текста на естественном языке основной проблемой является разрешение неоднозначностей. Так, в работе [7] было выявлено, что основной проблемой при частичном синтаксическом анализе в русском языке являются явления полисемии и омонимии, которые создают избыточность при обнаружении чанков. В настоящей работе предложено несколько эвристик, направленных на устранение избыточности.

1. Формальная постановка задачи

В предыдущих работах [7, 8] мы столкнулись с проблемой избыточности обнаружения чанков. Она определяется двумя явлениями в русском языке:

1) полисемия – наличие совпадающих лемм[1] в словаре. Например, *коса* на голове у девушки, *коса* как орудие труда и песчаная *коса* в море;

2) омонимия – наличие совпадающих морфоформ, то есть одно и то же написание слова соответствует большому количеству морфоформ слов с различными грамматическими характеристиками. Например, стали (Сущ., ж.р., Р.П., мн.ч.) и стали (Глаг., прош. вр., I спр., 3 лицо); или железа (Сущ., ср.р., Р.П., мн.ч.) и железа (Сущ., ж.р., И.П., ед.ч.).

Стоит отметить, что в нашем исследовании задача чанкинга решается в ограниченном масштабе, так как мы выявляем чанки только с именами существительными (ИС). Это мотивировано дальнейшим использованием результатов чанкинга для построения лексико-синтаксических портретов существительных [9], которые далее используются для разрешения многозначности ИС. Изначально в работе [8] была предложена эвристика, которая ограничивает дистанцию от ИС до зависимого слова. Эта эвристика улучшила результаты анализа на 32%. Прямой перебор без эвристик давал значение меры F1, равное 0.25, эвристика ограничения расстояния до зависимого слова позволила получить значение F1 равным 0.32 (надо заметить, что похожее исследование для чешского языка было произведено в 2005 г. [10]).

В настоящей работе мы предлагаем десять новых эвристик, направленных на уменьшение количества ложно выявленных чанков. При разработке этих эвристик мы предложили эффективные методы, которые позволили сократить количество неправильно выявленных чанков еще на 67% по сравнению с вариантом, который использовал только эвристику, ограничивающую дистанцию до зависимого слова. Нам удалось, используя все более и более глубокий анализ каждого предложения, избавиться от все большего количества ложно выявленных чанков.

Разделим эвристики на две группы:

А) эвристики, которые в алгоритме частичного синтаксического анализатора должны выполняться на этапе построения набора чанков для каждого предложения;

Б) эвристики, которые в алгоритме частичного синтаксического анализатора должны выполняться после этапа построения предварительного набора чанков для каждого предложения, то есть из числа уже имеющихся чанков исключаются ложные чанки.

[1]Лемма – нормализованная основная форма слова вместе с информацией о построении других форм.

Первая группа включает в себя следующие эвристики:

Эвристика А.1. Неверными являются чанки вида «Любое существительное (главное в чанке) + Существительное в именительном падеже».

Пример:

Предложение «Газ для доменной печи». Чанк «газ + печи» (Сущ., И.П. + Сущ., И.П.) неверный.

Эвристика А.2. Неверным считается чанк, где перед существительным в именительном падеже стоит любой предлог. Между существительным и предлогом также может стоять прилагательное.

Пример:

Предложение «В сталь добавляют молибден».

Чанк «сталь + добавляют» (Сущ., И.П. + Глагол) неверный.

Эвристика А.3. В основу этой эвристики легло одно из правил согласования предлогов с именами существительными в русском языке, а именно тот факт, что, судя по предлогу, стоящему перед существительным, можно сделать вывод, в каком падеже может стоять это существительное. Эвристика работает следующим образом: для каждого анализируемого существительного (в составе чанка любого вида) программа пытается найти предлог, который относится к этому существительному; если предлог найден, то необходимо проанализировать, в каком падеже может стоять найденное существительное. Если существительное стоит в таком падеже, в котором оно не может стоять в паре с предлогом, то программный комплекс делает вывод, что найденный чанк является ложным.

Пример:

Предложение «В ходе процесса из руды получается чугун».

Чанк «ходе + руды» (Сущ., Д.П. + Сущ., Р.П.) неверный.

Слово «ходе» может стоять только в винительном или предложном падежах, так как перед ним стоит предлог «в».

Ниже приведен список предлогов и падежей, в которых могут стоять существительные, идущие после предлогов.

Предлоги, которые можно употреблять с существительным, имеющим один падеж:

– с родительным: без, до, из, от, у, для, ради;
– с дательным: к;
– с винительным: про, через, сквозь;
– с творительным: над, перед;
– с предложным: при.

Предлоги, которые можно употреблять с существительным, имеющим один из двух падежей:

– с винительным и предложным: в, на, о;
– с родительным и творительным: между;
– с винительным и творительным: за, под.

Предлоги, которые можно употреблять с существительным, имеющим один из трех падежей:

– с винительным, дательным, предложным: по;
– с родительным, винительным и творительным: с.

Эвристика А.4. Данная эвристика показывает хорошие результаты для удаления ложных чанков, но в то же время она является «опровергаемой» эвристикой, то есть существуют предложения, в которых эта эвристика будет удалять также и верно найденные чанки.

Идея эвристики заключается в ограничении, накладываемом на чанки вида «Существительное + Существительное». Согласно данной эвристики чанк из двух

существительных считается неверным, если между ними стоит еще одно существительное.

Пример правильной работы эвристики (таких примеров очень много):

«Здесь содержится железо в виде соли».

Правильный чанк – «железо + в + виде», а неверный чанк «железо + соли» отсеивается.

Пример неверной работы эвристики:

«Повышение в чугуне количества соли».

«Повышение + в + чугуне» – правильный чанк, а вот чанк «повышение + количества» отсеивается неверно.

Данная эвристика не гарантирует стопроцентной точности. Однако решение об ее использовании было принято из-за того, что соотношение неверно удаленных правильных чанков к правильно удаленным ложным чанкам составляет около 1/11.

Далее рассмотрим вторую группу эвристик. Отметим тот факт, что данные эвристики не могут применяться на этапе построения набора чанков, так как исходными данными для этих эвристик являются все чанки, выявленные в предложении, то есть принять решение о том, является тот или иной чанк ложным, можно только после того, как стали известны все результаты работы эвристик из группы А.

Вторая группа (Б) включает в себя следующие эвристики:

Эвристика Б.1. Направлена на удаление из набора обнаруженных в предложении чанков тех, которые являются полисемичными к друг другу за исключением первого из них. Речь идет о том, что не важно, в каком значении в чанке присутствует каждое слово, главное – как структура чанка не содержит информацию о значении слов. Таким образом, нам удастся избавиться от большого количества чанков, которые не отличаются ни словами, которые составляют этот чанк, ни морфоформами этих слов.

Пример:

Предложение «У девочки есть коса».

Чанков «коса + есть» (Сущ., И.П. + Глагол) может быть 3 штуки, в зависимости от того, в каком смысле употребляется слово *коса*. Но на данном этапе анализа для нас не важно, в каком значении употребляется слово *коса*. Результатом же работы всей поисковой системы должно явиться определение, в каком именно значении употребляется слово *коса*.

Эвристика Б.2. Направлена на удаление чанков типа «Существительное + Существительное» в том случае, если в предложении между этими словами стоит союз «и», так как в этом случае эти два слова являются однородными членами или относятся к двум разным однородным предложениям, разделенным союзом «и» без запятой перед ним.

Пример:

Предложение «Молибден и хром добавляют в сталь».

Чанк «молибден + хром» неверный.

Эвристика Б.3. На первом этапе осуществляется поиск таких существительных, которые стоят в чанке в именительном падеже. Далее производится поиск по всем чанкам в том же простом предложении, которые содержат эти же слова, что и первоначально найденный чанк. Если таких чанков в исследуемом предложении не найдено или найденные чанки различаются формой слова, которое не является существительным, а существительное в этом случае во всех таких чанках находится в именительном падеже, это означает, что в простом предложении существует единственный кандидат на подлежащее. В этом случае все остальные

чанки в рамках простого предложения, в состав которых входят другие существительные в именительном падеже, точно являются неправильными и могут быть удалены из числа правильных чанков. Из этого правила существует исключение, хотя оно и маловероятно. Это исключение связано с тем, что, возможно, существуют два однородных предложения разделенных союзом «и» без запятой перед ним, в которых оба подлежащих подчиняются вышеописанным правилам, то есть оба этих подлежащих входят в чанки, для которых не существует омонимичных существительных не в именительном падеже. В этом случае возникает вопрос, какой из чанков оставлять, так как программный комплекс считает эти два однородных предложения одним простым предложением, что обусловлено тем, что в настоящем исследовании мы не разрабатываем и не применяем сложных методов разбиения сложного предложения на простые. Нами было решено, что в данном случае из этих двух «однородных» чанков оба будут оставаться в составе правильных чанков, что соответствует действительности.

Пример:

Предложение «По мере развития техники производства железа постепенно повышалась температура, при которой велся процесс».

Чанки «производства + железа (в смысле материал)» (Сущ., мн.ч., И.П. + Сущ., мн.ч., Р.П.), «железа (в смысле, орган) + повышалась» (Сущ., ед.ч., И.П. + Глагол) и т. д. неверные, так как в предложении только слово «температура» в чанке «температура + повышалась» стоит в именительном падеже.

Эвристика Б.4. Эвристика работает следующим образом. Если в простом предложении есть подлежащее (на данном этапе подлежащее определяется только при наличии сказуемого), то остальные существительные в этом предложении могут стоять в любом падеже, кроме именительного, а, соответственно, это слово – только в именительном падеже.

Пример:

Предложение «Зародился двухступенчатый способ производства железа из руды».

Чанк «зародился + способ» (Глагол + Сущ., В.П.) неверный, так как слово «способ» является подлежащим в данном предложении.

Эвристика Б.5. Идею пятой эвристики опишем по шагам (эти действия выполняются для каждого существительного в каждом чанке в простом предложении, в дальнейшем это существительное будем называть «эталонным существительным», а чанк, в который входит это существительное, – «эталонным чанком»):

Шаг 1. Производится проверка наличия хотя бы еще одного чанка «двойника», в который входят оба слова из «эталонного чанка», но «эталонное существительное» имеет другие морфологические характеристики, чем существительное из чанка «двойника». Шаг 2 выполняется только в случае положительного результата.

Шаг 2. Производится проверка наличия хотя бы еще одного чанка «полудвойника», в который входит это же «эталонное существительное», а остальные слова в чанке отличаются от слов из «эталонного чанка». Шаг 3 выполняется только в случае положительного результата.

Шаг 3. В том случае, если среди всех существительных из чанков «полудвойников» нет существительного с морфологическими характеристиками, аналогичными «эталонному существительному», а среди существительных из чанков «двойников» имеется хотя бы одно существительное, соответствующее по морфологическим характеристикам хотя бы одному существительному из чанков «полудвойников», «эталонный чанк» считается некачественным и удаляется из числа правильных чанков.

Пример:

Предложение «Температура стали по сравнению с чугуном больше».

Чанк «стали + по + сравнению» (Глаг., соверш. вида, невозврат., I спряж., прош. вр., мн. всех лиц родов) + Предлог + Сущ., Д.П.) неверный, так как во всех остальных чанках слово «стали» это существительное, но не глагол.

Эвристика Б.6. Эвристика, которая позволяет распознавать чанки, включающие составные глаголы. Во-первых, она позволяет сократить число ложных чанков, так как удаляет чанки, в состав которых входит существительное и один из двух глаголов. Во-вторых, она увеличивает количество верных чанков.

Пример:

Предложение «Чугун могут нагревать».

Чанк «чугун + могут нагревать» верный.

Эвристика Б.7. Была реализована также одна эвристика, от применения которой мы пока отказались (в программном комплексе ее можно включать и отключать, ее нет в математической постановке задачи), так как чтобы она работала правильно, ее необходимо сильно доработать. Она основывалась на том факте, что в большинстве предложений подлежащее предшествует сказуемому. Но поскольку это утверждение выполняется для большинства, но не для всех предложений, и более того, в некоторых текстах, таких, как стихи, это утверждение может быть вообще неверным, использовать данную эвристику без дополнительных правил невозможно.

2. Математическая постановка задачи

Рассмотрим математическую постановку задачи.

а) Предложение можно представить в виде упорядоченного множества слов

$$S = \{w_1, w_2, w_3, \ldots, w_n\}$$

и заданного на этом множестве отношения порядка

$$N_1 < N_2 < N_3 < \cdots < N_n, \tag{1}$$

где N_i – место слова w_i в предложении.

б) Этап морфоанализа можно представить в следующем виде:

$$(w_i^0, G_i) = MA(w_i),$$

где w_i^0 – нормальная форма слова; G_i – кортеж грамматических характеристик: $G_i = (g_{1_i}, g_{2_i}, \ldots, g_{n_i})$; $MA(w_i)$ – функция морфоанализа.

в) Теперь предложение может быть представлено в виде множества пар

$$T = \{(w_1^0, G_1), (w_2^0, G_2), \ldots, (w_n^0, G_n)\}$$

и заданного на этом множестве отношения порядка (1).

г) Расстояние между словами в предложении определяется как

$$Z = |i - j|$$

где i, j – позиции слов в предложении.

д) Поиск чанка (связанного словосочетания) сводится к перебору всех комбинаций пар в предложении и проверке выполнения условий

$$\underset{Z \le Z_0}{\mathrm{Comp}}\,(G_i, G_j) = \begin{cases} \text{True}, & \text{если } (A_{ij} = \text{True}) \wedge (B_{ij} = \text{True}) \wedge (C_{ij} = \text{True}), \\ \text{False}, & \text{если } (A_{ij} = \text{False}) \vee (B_{ij} = \text{False}) \vee (C_{ij} = \text{False}), \end{cases}$$

где $\mathrm{Comp}\,(G_i, G_j)$ – логическая функция сравнения; Z_0 – область поиска чанков в предложении; A_{ij} – условия для первоначального поиска чанков; B_{ij} – условия, описывающие эвристики, работающие на этапе поиска каждого чанка; C_{ij} – условия, описывающие эвристики, работающие после окончания поиска всех чанков в предложении.

е) $B_{ij} = (b_{1\,i,j}, b_{2\,i,j}, b_{3\,i,j}, b_{4\,i,j})$ – условия, описывающие эвристики, работающие на этапе поиска каждого чанка: $B_{ij} = \mathrm{True}$, если $b_{k\,i,j} = \mathrm{True},\ k = 1, 2, 3, 4$.

Пусть изначально для каждого предложения:

$$b_{k\,i,j} = \mathrm{True}, \quad i, j = n, \quad k = 1, 2, 3, 4.$$

А.1. Чанк неверный, если он имеет вид «Любое существительное (главное в чанке) + Существительное в именительном падеже»:

$$\left\{ \begin{array}{c} G_i = \text{Имя\_Сущ.}; \\ w_i - \text{главное\_слово\_в\_чанке}\ ; \\ G_j = \text{Имя\_Сущ. И.П.}; \end{array} \right\} \Rightarrow b_{1\,i,j} = \mathrm{False}.$$

А.2. Чанк неверный, если перед существительным в именительном падеже стоит любой предлог:

$$\left[\left\{ \begin{array}{c} G_i = \text{Имя\_Сущ. И.П.}; \\ i \geq 2; \\ G_{i-1} - \text{предлог}; \\ \text{Количество\_слов\_в\_предложении} \geq 3; \end{array} \right. \right.$$
$$\left. \left. \begin{array}{c} G_i = \text{Имя\_Сущ. И.П.}; \\ i \geq 3; \\ G_{i-1} - \text{прилагательное}; \\ G_{i-2} - \text{предлог}; \\ \text{Количество\_слов\_в\_предложении} \geq 4; \end{array} \right\} \right. \Rightarrow b_{2\,i,j} = \mathrm{False}.$$

А.3. Судя по предлогу, стоящему перед существительным, можно сделать вывод, в каком падеже может стоять это существительное [11].

Данная эвристика подробно описана в Приложении № 1.

А.4. Чанк из двух существительных считается неверным, если между ними стоит еще одно существительное (опровергаемая эвристика):

$$\left\{ \begin{array}{c} G_i = \text{Имя\_Сущ.} \\ G_j = \text{Имя\_Сущ.} \\ \left[\begin{array}{l} ((i < j) \wedge (G_{i+1} \neq \text{Имя\_Сущ.} \wedge \cdots \wedge G_{j-1} \neq \text{Имя\_Сущ.})) \\ ((i > j) \wedge (G_{j+1} \neq \text{Имя\_Сущ.} \wedge \cdots \wedge G_{i-1} \neq \text{Имя\_Сущ.})) \end{array} \right. \end{array} \right\} \Rightarrow b_{4\,i,j} = \mathrm{False}.$$

ж) $C_{ij} = (c_{1\,i,j}, c_{2\,i,j}, c_{3\,i,j}, c_{4\,i,j}, c_{5\,i,j}, c_{6\,i,j})$ – условия, описывающие эвристики, работающие после окончания поиска всех чанков в предложении:

$$C_{ij} = \mathrm{True}, \quad \text{если } c_{k\,i,j} = \mathrm{True}, \quad k = 1, 2, \ldots, 6.$$

Пусть изначально для каждого предложения:

$$c_{k\,i,j} = \mathrm{True}, \quad i, j = n, \quad k = 1, 2, \ldots, 6.$$

Б.1. Удаление из набора обнаруженных в предложении чанков тех, которые являются полисемичными к друг другу, за исключением первого из них:

$$c_{1\,i,j} = \mathrm{True}, \quad \text{если } (w_i^0, G_i) \neq (w_j^0, G_j);$$

$$c_{1\,i,j} = \text{False}, \quad \text{если } (w_i^0, G_i) = (w_j^0, G_j).$$

Б.2. Эвристика направлена на удаление чанков типа «Существительное + Существительное» в том случае, если в предложении между этими словами стоит союз «и»:

$$c_{2\,i,j} = \text{True}, \quad \text{если} \left[\begin{array}{c} ((i<j) \wedge (w_{i+1}^0 \neq \text{“и”} \wedge \cdots \wedge w_{j-1}^0 \neq \text{“и”})) \\ ((i>j) \wedge (w_{j+1}^0 \neq \text{“и”} \wedge \cdots \wedge w_{i-1}^0 \neq \text{“и”})) \end{array} \right. ;$$

$$c_{2\,i,j} = \text{False}, \quad \text{если} \left[\begin{array}{c} ((i<j) \wedge (w_{i+1}^0 = \text{“и”} \vee \cdots \vee w_{j-1}^0 = \text{“и”})) \\ ((i>j) \wedge (w_{j+1}^0 = \text{“и”} \vee \cdots \vee w_{i-1}^0 = \text{“и”})) \end{array} \right. .$$

Б.3. Если в простом предложении существует только одно существительное в именительном падеже, то это существительное может присутствовать в составе чанков только в именительном падеже, а все остальные существительные должны стоять не в именительных падежах:

$$\left[\begin{array}{c} \left\{ \begin{array}{c} (G_i = \text{Имя\_Сущ. И.П.}) \\ (w_i^0 \neq \forall w_{1..n}^0) \wedge (G_i \neq \forall G_{1..n}) \\ \neg\exists((i \neq h) \wedge (w_c^0 \neq \forall w_{1..n}^0) \wedge (G_c = \text{Имя\_Сущ. И.П.})) \\ ((w_i^0 \neq w_k^0) \wedge (G_k = \text{Имя\_Сущ. И.П.})) \end{array} \right\} \\ \left\{ \begin{array}{c} (G_j = \text{Имя\_Сущ. И.П.}) \\ (w_j^0 \neq \forall w_{1..n}^0) \wedge (G_j \neq \forall G_{1..n}) \\ \neg\exists((j \neq h) \wedge (w_c^0 \neq \forall w_{1..n}^0) \wedge (G_c = \text{Имя\_Сущ. И.П.})) \\ ((w_j^0 \neq w_k^0) \wedge (G_k = \text{Имя\_Сущ. И.П.})) \end{array} \right\} \end{array} \right. \Rightarrow c_{3k,1\ldots n} = \text{False}.$$

Б.4. Если в простом предложении есть подлежащее (на данном этапе подлежащее определяется только при наличии сказуемого), то остальные существительные в этом предложении могут стоять в любом падеже, кроме именительного, а, соответственно, это слово – только в именительном падеже:

$$\left\{ \begin{array}{c} \exists(w_i^0, G_i) = \text{Подлежащее} \\ \left[\begin{array}{c} (G_i \neq \text{Имя\_Сущ. И.П.}) \Rightarrow x = i \\ (i \neq h) \wedge (G_h = \text{Имя\_Сущ. И.П.}) \Rightarrow x = h \end{array} \right] \end{array} \right\} \Rightarrow c_{4x,j} = \text{False}.$$

Б.5. Эвристика, определяющая составные глаголы в предложении.

Данная эвристика подробно описана в Приложении №2.

Б.6. Эвристика, проверяющая «качество» чанков в предложении.

Эта эвристика подробно описана в Приложении № 3.

Если $\text{Comp}\,(G_i, G_j) = \text{True}$, то пара слов (w_i, w_j) является чанком, в противном случае пара слов (w_i, w_j) не является чанком.

Для тестирования данного алгоритма был создан специальный программный комплекс «Chunk-Creator 4» [7, 8].

3. Анализ эффективности применяемой эвристики

Для анализа эффективности применяемой эвристики необходимо рассчитать:

N_1 – число правильно выявленных чанков, то есть количество истинно выявленных чанков в тестируемой выборке (мнение эксперта совпало с результатом автоматической классификации, чанки были определены и экспертом, и программным комплексом);

N_2 – число неправильно выявленных чанков, а именно количество ложно приписанных чанков в тестируемой выборке (эксперт не обнаружил чанк, а программный комплекс обнаружил);

N_3 – число пропущенных чанков, то есть число ложно отклоненных чанков в тестируемой выборке (эксперт обнаружил чанк, а программный комплекс не обнаружил).

Эффективность работы алгоритма оценивается следующим образом.

Рассчитывается P_r – показатель точности классификации, которая рассчитывается как отношение числа правильно выявленных чанков к сумме числа правильно выявленных чанков и числа пропущенных чанков:

$$P_r = \frac{N_p}{N_p + N_n},$$

Рассчитывается R_e – показатель отказа классификации, которая рассчитывается как отношение числа правильно выявленных чанков к сумме числа правильно выявленных чанков и числа неправильно выявленных чанков:

$$R_e = \frac{N_p}{N_p + N^*}.$$

Далее рассчитывается результирующая мера F_1 по формуле

$$F_1 = \frac{2 \cdot P_r \cdot R_e}{P_r + R_e}.$$

В результате тестирования программного комплекса были получены следующие данные (табл. 1).

В табл. 2 приведены результаты, полученные без применения эвристик, описываемых в настоящей статье, но с применением эвристики, накладывающей ограничение на расстояние до зависимого слова, которая описана в статье [8].

Здесь для примера приведен анализ части одного предложения с применением эвристик с областью анализа ± 3 слова (табл. 3).

В нашем примере видно, что произошло сокращение неправильно выявленных чанков с 99 до 24. При этом не было потеряно ни одного правильного чанка.

Теперь проанализируем полученные результаты из табл. 3, где представлены общие результаты анализа.

Видно, что ложных чанков осталось намного меньше, чем было раньше. Большинство оставшихся ложных чанков не удалось удалить вследствие того, что не удалось определить падеж каждого существительного со стопроцентной вероятностью.

В чанках вида «Глагол + Существительное» также не применяется эвристика, ограничивающая область анализа, так как многие правильные чанки именно этого вида могут неверно удаляться данной эвристикой. Поэтому у нас есть 2 новых ложных чанка «повышалось + углерода» и «повышалось + примесей».

Далее проанализируем данные, представленные в табл. 1 и 2.

Обратим внимание на то, что эти эвристики обладают преимуществом перед эвристикой, направленной на ограничение области анализа зависимого слова из-за того, что они никак не воздействуют на правильно выявленные морфоанализатором чанки (за исключением эвристики 1.4).

Отметим, что значение P_r возросло с 0.90 до 0.94 (при области анализа в 3 слова). Значение R_e увеличилось для каждой области анализа в среднем на 0.22–0.26, а F_1 – на 0.28.

Можно также отметить, что наилучшие результаты дает анализ с областью ± 2, ± 3 и ± 4 слова, при этом ошибочно отсеивается соответственно 8.9%, 2.76% и 0.61% чанков. Число неправильно выявленных чанков вследствие применения эвристик

Табл. 1

Анализ эффективности применяемых эвристик

	Область анализа (± количество слов от существительного)							
	Все слова	7	6	5	4	3	2	1
N_1	357	357	357	357	355	347	325	250
N_2	628	597	581	552	503	432	344	201
N_3	16	16	16	16	17	22	35	80
$\sum_{\text{прог}}$	985	954	937	909	858	779	669	450
$\sum_{\text{вруч}}$	373	373	373	373	372	369	360	329
$\%_{\text{прав}}$	95.66	95.66	95.66	95.66	95.31	94.04	90.30	75.82
$\%_{\text{проп}}$	4.34	4.34	4.34	4.34	4.69	5.96	9.70	24.18
Усл. $\%_{\text{непр}}$	168.35	160.09	155.65	148.00	135.25	117.07	95.45	60.94
Потеря $_{\text{п.ч.}}$	0.00	0.00	0.00	0.00	0.61	2.76	8.90	30.06
Потеря $_{\text{н.ч.}}$	0.00	4.91	7.55	12.09	19.85	31.21	45.27	68.06
P_r	0.96	0.96	0.96	0.96	0.95	0.94	0.90	0.76
R_e	0.36	0.37	0.38	0.39	0.41	0.45	0.49	0.55
F_1	0.53	0.54	0.54	0.56	0.58	0.60	0.63	0.64

где $\sum_{\text{прог}} = N_1 + N_2$ – всего чанков (определил чанкер);

$\sum_{\text{вруч}} = N_1 + N_3$ – всего чанков (посчитано вручную);

$\%_{\text{прав}} = N_1 / \sum_{\text{вруч}} \cdot 100\%$ – процент правильно выявленных от $\sum_{\text{вруч}}$;

$\%_{\text{проп}} = N_3 / \sum_{\text{вруч}} \cdot 100\%$ – процент пропущенных от $\sum_{\text{вруч}}$;

Усл. $\%_{\text{непр}} = N_2 / \sum_{\text{вруч}} \cdot 100\%$ – условный процент неправильно выявленных чанков от $\sum_{\text{вруч}}$;

$\text{Потеря}_{\text{п.ч.}} = 100 - (N_1 / N_{1\_\text{ВСЕ\_СЛОВА}} \cdot 100\%)$ – процент потерянных правильных чанков вследствие применения эвристики;

$\text{Потеря}_{\text{н.ч.}} = 100 - (N_3 / N_{1\_\text{ВСЕ\_СЛОВА}} \cdot 100\%)$ – процент отсеянных неправильных чанков вследствие применения эвристики.

Табл. 2

Анализ эффективности с применением эвристики, связанной с ограничением расстояния до зависимого слова, и без применения эвристик, описываемых в настоящей статье

	Область анализа (± количество слов от существительного)							
	Все слова	7	6	5	4	3	2	1
N_1	326	326	326	326	324	317	297	228
N_2	1923	1822	1761	1687	1555	1316	1002	585
N_3	25	25	25	25	27	34	54	123
$\sum_{\text{прог}}$	2249	2148	2087	2013	1879	1633	1299	813
$\sum_{\text{вруч}}$	351	351	351	351	351	351	351	351
$\%_{\text{прав}}$	92.88	92.88	92.88	92.88	92.31	90.31	84.62	64.96
$\%_{\text{проп}}$	7.12	7.12	7.12	7.12	7.69	9.69	15.38	35.04
Усл. $\%_{\text{непр}}$	547.86	519.09	501.71	480.63	443.02	374.93	285.47	166.67
Потеря $_{\text{п.ч.}}$	0.00	0.00	0.00	0.00	0.61	2.76	8.90	30.06
Потеря $_{\text{н.ч.}}$	0.00	5.25	8.42	12.27	19.14	31.57	47.89	69.58
P_r	0.93	0.93	0.93	0.93	0.92	0.90	0.85	0.65
R_e	0.14	0.15	0.16	0.16	0.17	0.19	0.23	0.28
F_1	0.25	0.26	0.27	0.28	0.29	0.32	0.36	0.39

Табл. 3

Поиск чанков в предложении с использованием эвристик

Цепь	Вид цепи	N_1	N_2	N_3	N_4	П
Предложение: «По мере развития техники производства железа постепенно повышалась температура, при которой велся процесс, но одновременно в металле повышалось содержание углерода и других примесей».						
повышалась + по + мере	Глаг. НСВ невозврат I спряж (Прош.вр Ед.всех лиц Ж род) + Предлог + Сущ. Ж.род. Неодуш. (Ед.ч. Д.П.)	1				1
мере + развития	Сущ. Ж.род. Неодуш. (Ед.ч. Д.П.) + Сущ. Ср.род. Неодуш. (Ед.ч. Р.П.)	1				1
мере + развития	Сущ. Ж.род. Неодуш. (Ед.ч. Д.П.) + Сущ. Ср.род. Неодуш. (Мн.ч. В.П.)		1			
повышалась + по + мере	Глаг. НСВ невозврат I спряж (Прош.вр Ед.всех лиц Ж род) + Предлог + Сущ. Ж.род. Неодуш. (Ед.ч. П.П.)		1			
мере + развития	Сущ. Ж.род. Неодуш. (Ед.ч. П.П.) + Сущ. Ср.род. Неодуш. (Ед.ч. Р.П.)		1			
мере + развития	Сущ. Ж.род. Неодуш. (Ед.ч. П.П.) + Сущ. Ср.род. Неодуш. (Мн.ч. В.П.)		1			
развития + техники	Сущ. Ср.род. Неодуш. (Ед.ч. Р.П.) + Сущ. Ж.род. Неодуш. (Ед.ч. Р.П.)	1				1
развития + техники	Сущ. Ср.род. Неодуш. (Ед.ч. Р.П.) + Сущ. Ж.род. Неодуш. (Мн.ч. В.П.)		1			
развития + техники	Сущ. Ср.род. Неодуш. (Мн.ч. В.П.) + Сущ. Ж.род. Неодуш. (Ед.ч. Р.П.)		1			
развития + техники	Сущ. Ср.род. Неодуш. (Мн.ч. В.П.) + Сущ. Ж.род. Неодуш. (Мн.ч. В.П.)		1			
техники + производства	Сущ. Ж.род. Неодуш. (Ед.ч. Р.П.) + Сущ. Ср.род. Неодуш. (Ед.ч. Р.П.)	1				1
техники + производства	Сущ. Ж.род. Неодуш. (Ед.ч. Р.П.) + Сущ. Ср.род. Неодуш. (Мн.ч. В.П.)		1			
техники + производства	Сущ. Ж.род. Неодуш. (Мн.ч. В.П.) + Сущ. Ср.род. Неодуш. (Ед.ч. Р.П.)		1			
техники + производства	Сущ. Ж.род. Неодуш. (Мн.ч. В.П.) + Сущ. Ср.род. Неодуш. (Мн.ч. В.П.)		1			
производства + железа	Сущ. Ср.род. Неодуш. (Ед.ч. Р.П.) + Сущ. Ср.род. Неодуш. (Ед.ч. Р.П.)	1				1
производства + железа	Сущ. Ср.род. Неодуш. (Ед.ч. Р.П.) + Сущ. Ср.род. Неодуш. (Мн.ч. В.П.)		1			
производства + железа	Сущ. Ср.род. Неодуш. (Мн.ч. В.П.) + Сущ. Ср.род. Неодуш. (Ед.ч. Р.П.)		1			
производства + железа	Сущ. Ср.род. Неодуш. (Мн.ч. В.П.) + Сущ. Ср.род. Неодуш. (Мн.ч. В.П.)		1			
температура + повышалась	Сущ. Ж.род. Неодуш. (Ед.ч. И.П.) + Глаг. НСВ невозврат I спряж (Прош.вр Ед.всех лиц Ж род)	1				1
повышалось + в + металле	Глаг. НСВ невозврат I спряж (Прош.вр Ед.всех лиц С род) + Предлог + Сущ. М.род. Неодуш. (Ед.ч. П.П.)	1				1
содержание + повышалось	Сущ. Ср.род. Неодуш. (Ед.ч. И.П.) + Глаг. НСВ невозврат I спряж (Прош.вр Ед.всех лиц С род)	1				1
содержание + углерода	Сущ. Ср.род. Неодуш. (Ед.ч. И.П.) + Сущ. М.род. Неодуш. (Ед.ч. Р.П.)	1				1
повышалось + углерода	Глаг. НСВ невозврат I спряж (Прош.вр Ед.всех лиц С род) + Сущ. М.род. Неодуш. (Ед.ч. Р.П.)		1			
повышалось + примесей	Глаг. НСВ невозврат I спряж (Прош.вр Ед.всех лиц С род) + Сущ. Ж.род. Неодуш. (Мн.ч. Р.П.)		1			
Итого		9	15	0	0	1
Примечание: «П» - правильные чанки в предложении, определяется вручную; «N_4» – неправильно выявленные чанки из-за ошибки морфоанализатора.						

соответственно для областей анализа ± 2, ± 3 и ± 4 равно: 51.42% (было 1002, стало 344 чанка), 55.46% (было 1316, стало 432 чанков) и 58.62% (было 1555, стало 503 чанков). В среднем количество неверно обнаруженных чанков сократилось втрое.

Так как потеря правильных чанков для областей анализа ± 2, ± 3 и ± 4 равна соответственно 8.90%, 2.76%, 0.61%, можно сделать вывод о том, что оптимальной областью анализа является область ± 3, при этой области потеря правильных чанков будет составлять 2.76%, $P_r = 0.94$, $R_e = 0.45$, $F_1 = 0.60$.

На рис. 1 представлены сравнительные графики значений P_r, R_e, F_1 в зависимости от области анализа для случая с применением эвристик. На рис. 2 представлены Сравнительные графики значений P_r, R_e, F_1 в зависимости от области анализа для случая без применения эвристик. На рис. 3 представлена сводная диаграмма значений P_r, R_e, F_1 для случаев с применением всех эвристик, описанных в настоящей статье, и для случая применения только одной эвристики, ограничивающей область анализа.

4. Выводы и направления будущих исследований

Применение данных эвристик не влияет на количество правильно выявленных чанков (за исключением эвристики 1.4). Число неправильно выявленных чанков

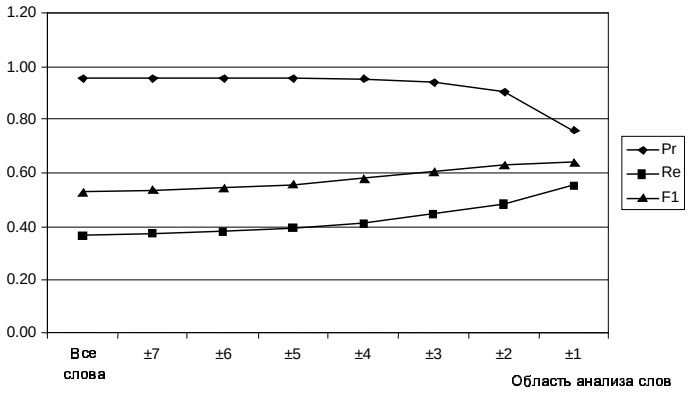

Рис. 1. Сравнительные графики значений P_r, R_e, F_1 в зависимости от области анализа для случая с применением всех эвристик

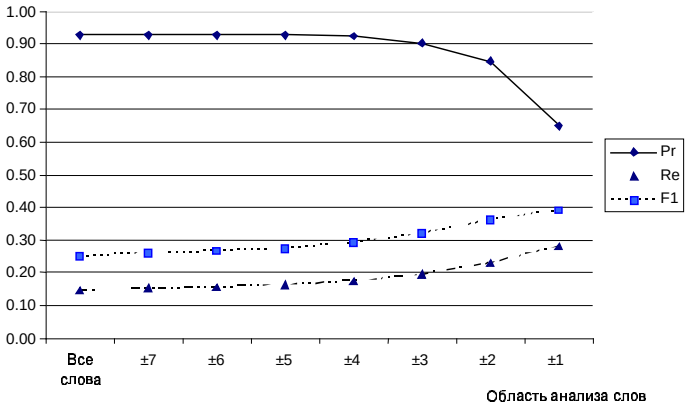

Рис. 2. Сравнительные графики значений P_r, R_e, F_1 в зависимости от области анализа для случая без применения эвристик

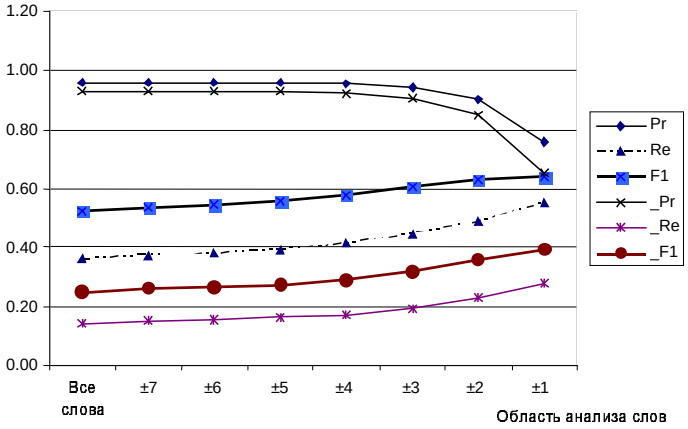

Рис. 3. Сводная диаграмма: P_r, R_e, F_1 – величины, полученные при применении всех эвристик, описанных в данной статье; $\_P_r$, $\_R_e$, $\_F_1$ – величины, полученные при применении только одной эвристики, ограничивающей область анализа

вследствие применения эвристик для разных областей анализа предложения примерно равно от 30% до 35% от первоначального количества (после применения эвристики, направленной на ограничение области анализа в предложении). При этом показатель R_e удалось увеличить на 0.22–0.27 (в абсолютном выражении).

Будущие исследования будут направлены на разработку новых синтаксических правил – эвристик, которые позволят еще больше сократить число ложно выявленных чанков, а также на исследование эвристик, описывающих дополнительные правила, позволяющие производить поиск новых видов чанков.

Приложение 1. Эвристика, которая по предлогу перед существительным, делает вывод о допустимости чанка:

$$
\left[\begin{cases} G_i \neq Имя\_Сущ.И.П.; \\ i \geq 2; \\ G_{i-1} - предлог; x = i-1; \\ Количество\_слов\_в\_предложении \geq 3; \end{cases}
\begin{cases} G_i \neq Имя\_Сущ.И.П.; \\ i \geq 3; \\ G_{i-1} - прилагательное; \\ G_{i-2} - предлог;; x = i-2; \\ Количество\_слов\_в\_предложении \geq 4; \end{cases} \Rightarrow \right.
$$

$$
\Rightarrow \left[\begin{cases} G_i = Имя\_Сущ.Р.П.; \\ w_x \neq "без" \wedge "до" \wedge "из" \wedge "от" \wedge "у" \wedge "для" \wedge "ради" \wedge "между" \wedge "с"; \end{cases}
\begin{cases} G_i = Имя\_Сущ.Д.П.; \\ w_x \neq "к" \wedge "по"; \end{cases}
\begin{cases} G_i = Имя\_Сущ.В.П.; \\ w_x \neq "про" \wedge "через" \wedge "сквозь" \wedge "в" \wedge "на" \wedge "о" \wedge "за" \wedge "над" \wedge "по" \wedge "с"; \end{cases} \Rightarrow
\begin{cases} G_i = Имя\_Сущ.Т.П.; \\ w_x \neq "над" \wedge "перед" \wedge "между" \wedge "за" \wedge "под" \wedge "с"; \end{cases}
\begin{cases} G_i = Имя\_Сущ.П.П.; \\ w_x \neq "при" \wedge "в" \wedge "на" \wedge "о" \wedge "по"; \end{cases} \right]
$$

$$
\Rightarrow b_{3i,j} = False.
$$

Приложение 2. Эвристика, определяющая составные глаголы в предложении:

$$
\left[\begin{cases} G_i \neq Имя\_Сущ.И.П.; \\ i \geq 2; \\ G_{i-1} - предлог; x = i-1; \\ Количество\_слов\_в\_предложении \geq 3; \end{cases}
\begin{cases} G_i \neq Имя\_Сущ.И.П.; \\ i \geq 3; \\ G_{i-1} - прилагательное; \\ G_{i-2} - предлог;; x = i-2; \\ Количество\_слов\_в\_предложении \geq 4; \end{cases} \Rightarrow \right.
$$

$$
\Rightarrow \left[\begin{cases} G_i = Имя\_Сущ.Р.П.; \\ w_x \neq "без" \wedge "до" \wedge "из" \wedge "от" \wedge "у" \wedge "для" \wedge "ради" \wedge "между" \wedge "с"; \end{cases}
\begin{cases} G_i = Имя\_Сущ.Д.П.; \\ w_x \neq "к" \wedge "по"; \end{cases}
\begin{cases} G_i = Имя\_Сущ.В.П.; \\ w_x \neq "про" \wedge "через" \wedge "сквозь" \wedge "в" \wedge "на" \wedge "о" \wedge "за" \wedge "над" \wedge "по" \wedge "с"; \end{cases} \Rightarrow
\begin{cases} G_i = Имя\_Сущ.Т.П.; \\ w_x \neq "над" \wedge "перед" \wedge "между" \wedge "за" \wedge "под" \wedge "с"; \end{cases}
\begin{cases} G_i = Имя\_Сущ.П.П.; \\ w_x \neq "при" \wedge "в" \wedge "на" \wedge "о" \wedge "по"; \end{cases} \right]
$$

$$
\Rightarrow b_{3i,j} = False.
$$

Приложение 3. Эвристика, проверяющая «качество» чанков в предложении:

$$\forall(w_{i"эталон"} + w_{j"эталон"}, G_{i"эталон"}, G_{j"эталон"}) - эталонный\_чанк$$

$$I.\left[\left\{\begin{array}{c} G_{i"эталон"} = Существительное; \\ \exists(w_{i"двойник"} + w_{j"двойник"}, G_{i"двойник"}, G_{j"двойник"}), \\ w_{i"эталон"} = w_{i"двойник"}, w_{j"эталон"} = w_{j"двойник"}, G_{i"эталон"} \neq G_{i"двойник"}; \\ G_{j"эталон"} = Существительное; \\ \exists(w_{j"двойник"} + w_{i"двойник"}, G_{j"двойник"}, G_{i"двойник"}), \\ w_{j"эталон"} = w_{j"двойник"}, w_{i"эталон"} = w_{i"двойник"}, G_{j"эталон"} \neq G_{i"двойник"}; \end{array}\right\}\right] \Rightarrow$$

$$II.\left[\left\{\begin{array}{c} G_{i"эталон"} = Существительное; \\ \exists(w_{i"полудв."} + w_{j"полудв."}, G_{i"полудв."}, G_{j"полудв."}), \\ w_{i"эталон"} = w_{i"полудв."}, w_{j"эталон"} \neq w_{j"полудв."}, G_{i"эталон"} = G_{i"полудв."}; \\ G_{j"эталон"} = Существительное; \\ \exists(w_{j"полудв."} + w_{i"полудв."}, G_{j"полудв."}, G_{i"полудв."}), \\ w_{j"эталон"} = w_{j"полудв."}, w_{i"эталон"} \neq w_{i"полудв."}, G_{j"эталон"} = G_{j"полудв."}; \end{array}\right\}\right] \Rightarrow$$

$$III.\left[\left\{\begin{array}{c} G_{i"эталон"} = Существительное; \\ \forall(w_{i"полудв."} + w_{j"полудв."}, G_{i"полудв."}, G_{j"полудв."}), \\ i"полудв."=1..кол-во\_полудв., j"полудв."=1..кол-во\_полудв.; \\ w_{i"эталон"} = w_{i"полудв."}; \\ G_{i"эталон"} \neq G_{i"полудв."}; \\ \forall(w_{i"двойник"} + w_{j"двойник"}, G_{i"двойник"}, G_{j"двойник."}), \\ i"двойник"=1..кол-во\_двойн., j"двойник"=1..кол-во\_двойников; \\ w_{i"двойник"} = w_{i"полудв."}; \\ G_{i"двойник"} = G_{i"полудв."}; \\ G_{j"эталон"} = Существительное; \\ \forall(w_{j"полудв."} + w_{i"полудв."}, G_{j"полудв."}, G_{i"полудв."}), \\ j"полудв."=1..кол-во\_полудв., i"полудв."=1..кол-во\_полудв.; \\ w_{j"эталон"} = w_{j"полудв."}; \\ G_{j"эталон"} \neq G_{j"полудв."}; \\ \forall(w_{j"двойник"} + w_{i"двойник"}, G_{j"двойник"}, G_{i"двойник."}), \\ j"двойник"=1..кол-во\_двойн., i"двойник"=1..кол-во\_двойников; \\ w_{j"двойник"} = w_{j"полудв."}; \\ G_{j"двойник"} = G_{j"полудв."}; \end{array}\right\}\right] \Rightarrow$$

$$\Rightarrow\left[\begin{array}{l} G_{i"эталон"} = Существительное \Rightarrow c_{5i"эталон",j"эталон"} = False \\ G_{j"эталон"} = Существительное \Rightarrow c_{5j"эталон",i"эталон"} = False \end{array}\right].$$

Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований (проект № 09-07-97007-р-поволжье_а).

Summary

V.A. Bushtedt, V.N. Polyakov. A Partial Parser with Heuristics Reducing the Number of False Chunks in the Russian Clause.

The problem of partial parsing is considered in this paper. New heuristics are proposed to reduce the quantity of chunks falsely exposed at the first step of analysis. A very large influence is rendered by the phenomena of homonymy and polysemy on detection of chunks in Russian. Falsely exposed chunks are treated as ones which were found out by a partial parser, but are not actually correct. The method of search of chunks with the use of these heuristics got the name "Right-chunk 4". The formal task statement is carried out. Computer realization of method of search of chunks is executed as software "Chunk-creator 4". The estimation of quality is conducted.

Key words: artificial intelligence, computational linguistics, parsing, chunking.

Литература

1. *Попов Э.В.* Общение с ЭВМ на естественном языке. – М.: Едиториал УРСС, 2004. – 360 с.

2. *Смирнов Ю.М., Андреев А.М., Березкин Д.В., Брик А.В.* Об одном способе построения синтаксического анализатора текстов на естественном языке // Изв. вузов. Приборостроение. – 1997. – Т. 40, № 5. – С. 34–42.

3. *Ермаков А.Е.* Неполный синтаксический анализ текста в информационно-поисковых системах // Компьютерная лингвистика и интеллектуальные технологии: Труды Междунар. семинара Диалог'2002: в 2 т. – М.: Наука, 2002, – Т. 2. – С. 180–185.

4. *Ермаков А.Е.* Тематический анализ текста с выявлением сверхфразовой структуры // Информ. технол. – 2000. – № 11. – С. 37–40.

5. *Ермаков А.Е., Плешко В.В.* Ассоциативная модель порождения текста в задаче классификации // Информ. технол. – 2000. – № 12. – С. 34–37.

6. *Андреев А.М., Березкин Д.В., Брик А.В., Кантонистов Ю.А.* Вероятностный синтаксический анализатор для информационно-поисковой системы // Компьютерная хроника. – 1999. – № 1. – С. 3–4.

7. *Буштедт В.А., Поляков В.Н.* Частичный синтаксический анализатор для корпоративной поисковой системы // Труды Казан. школы по компьютерной и когнитивной лингвистике (TEL-2006). – Казань: Отечество, 2007. – С. 4–15.

8. *Bushtedt V., Polyakov V.* Finding chunks with restricrion of distance to dependent word // Когнитивное моделирование в лингвистике: Труды IX междунар. конф. – Sofia, Bulgaria, 2007. – С. 38–47.

9. *Кузьмин Ю.Г., Поляков В.Н., Шмагина Е.В.* Метод лексико-синтаксических портретов и задача разрешения лексической многозначности // Труды Казан. школы по компьютерной и когнитивной лингвистике (TEL-2006). – Казань: Отечество, 2007. – С. 139–147.

10. *Hall K., Novak V.* Corrective modeling for non-projective dependency parsing // CProceedings of the 9th International Workshop on Parsing Technologies (IWPT). – 2005. – P. 42–52.

11. Современный русский язык: Лексика и фразеология. Фонетика и орфоэпия. Графика и орфография. Словообразование. Морфология. Синтаксис: Учебник для вузов / Под ред. Д.Э. Розенталя. – М.: Высш. шк., 1984. – 735 с.

Поступила в редакцию
26.02.09

Поляков Владимир Николаевич – кандидат технических наук, доцент Московского государственного лингвистического университета и Московского института стали и сплавов, старший научный сотрудник Института языкознания РАН, г. Москва.

E-mail: pvn-65@mail.ru

Буштедт Владислав Андреевич – аспирант Московского института стали и сплавов.

E-mail: chap_007@mail.ru