ПОВТОРНОЕ ИСПОЛЬЗОВАНИЕ СЛОВАРЕЙ ПРИ СОЗДАНИИ НОВЫХ

В.В.Смирнов

Московский государственный инженерно-физический институт (технический университет),
115409 Москва, Каширское шоссе, 31, МИФИ, Vitaly_Smirnov@mail.ru

В работе рассматривается опыт атоматизации извлечения лексики из текстов для формирования новых словарей, используя словарь базовой лексики, статистику встречаемости и сочетаемости слов.

Задача повторного использования существующих базовых словарей для создания новых часто возникает при развитии готовых или создании новых систем обработки текста. Аналогичная задача возникла при наполнения словарей специализированного лингвистического процессора (ЛП) инструментального комплекса АТ-ТЕХНОЛОГИЯ[Колобашкина и др., 1996; Рыбина и др., 1997, Рыбина и др., 2001], преобразующего фразы ограниченного естественного языка (ЕЯ) в предикатно-аргументные структуры на языке CAREL[Рыбина, 1990].

Инструментальный комплекс АТ-ТЕХНОЛОГИЯ предназначен для автоматизированного построения прикладных интегрированных экспертных систем (ИЭС) в статических проблемных областях на основе использования задачно-ориентированной методологии (ЗОМ) компьютерного построения ИЭС[Рыбина, 1997].

ЛП в сочетании с другими программными средствами (ПС) комплекса АТ-ТЕХНОЛОГИЯ обеспечивает решение целого ряда задач, в частности:

- определение типа решаемой задачи [Пышагин и др., 1996] для активизации соответствующего ему сценария диалога с экспертом;
- извлечение лексики системного аналитика (СА) [Колобашкина и др., 1996; Рыбина и др., 1997; Рыбина и др., 2001] и эксперта из текста с описанием решаемой задачи [Рыбина и др., 1998] и протокола интервьюирования эксперта (ПИЭ) [Рыбина и др., 1997] с формированием вспомогательных словарей новой лексики;
- получение информации о наличии НЕ-факторов [Рыбина и др., 2001; Рыбина, Душкин и др., 2001] (например, определение ограничений на значения атрибутов базы знаний, значений лингвистических переменных и другой информации о НЕ-факторах в ответах на вопросы, полученных в результате интервьюирования эксперта).

Определение типа решаемой задачи выполняется при извлечении знаний методом «имитации консультации»[Рыбина и др., 1999; Кустикова и др., 1998] в начале сеанса интервьюирования эксперта на основе анализа текста, в котором эксперт описывает решаемую задачу.

Например, при описании решаемой задачи эксперт может ввести следующую фразу «Выявить у пациента интерстициальный фиброз легких» которая будет представлена в виде следующего CAREL-выражения:

(выявить) (МОД (INFN))
(D (пациент) (Оду))
(O (фиброз) (Заб)

(H (интерстициальный) (Хар))

(H (легкое) (Нео)))

где, «выявить» - предикат; D - глубинный падеж «быть адресатом действия»; O - глубинный падеж «быть объектом действия»; H – обозначение связи типа «быть характеристикой»; МОД, INFN - зарезервированные слова для обозначения модальности; Оду, Заб, Хар, Нео - обозначения типов семантических категорий в принятой кодировке использованного в данном случае словаря.

Такое представление фразы в виде данного CAREL-выражениея позволяет ЛП отнести описанную экспертом задачу к типу «Диагностика».

В процессе извлечения лексики формируются новые словари, дополняющие словари базовой лексики ЛП. При наполнении новых словарей возник ряд проблем, при решении которых в структуру словарей были внесены изменения, а для обработки текста использованы дополнительные знания о его структуре.

Обработка ЕЯ с помощью ЛП выполняется при помощи морфологического, синтаксического и семантического анализа с использованием словарей квазифлексий, предикатов, понятий, характеристик и классификатора семантических классов и семантических категорий. Для морфологического анализа ЛП использует процедурный метод [Андреев и др., 1998], который предполагает использование словарей основ со ссылкой на строки в таблице возможных аффиксов. В первых версиях ЛП словари для хранения предикатов и характеристик также были организованы по принципу разделения основ и аффиксов, однако это существенно усложняло алгоритм формирования новых словарей, так как вынуждало заносить слова, имеющие изменяющуюся основу, в словарь исключений. Использование в последних версиях ЛП словарей квазиоснов и квазифлексий [Автоматизация анализа научного текста, 1984], позволило отказаться от применения словаря исключений. Использование словаря квазифлексий позволило также во многих случаях решить проблему автоматического определения морфологических характеристик новых слов, в частности установления их
канонической формы, к которой приводятся все слова обрабатываемой ЛП фразы при формировании предикатно-аргументной структуры.

У каждой квазиосновы слова в словаре квазиоснов имеется ссылка на строку в таблице квазифлексий, соответствующей определенному типу словоизменения, которая может быть пустой, в случае если слово не имеет форм (например, аббревиатуры, сокращения, частицы, предлоги и т.д.) или если другие формы пока не известны. При автоматическом определении морфологических характеристик новых слов выполняется сравнение новых слов с таблицей квазифлексий, которую, можно рассматривать как расширение морфологической таблицы русского языка за счет включения туда частей корня слов. При выполнении сравнения используется дополнительная индексация таблицы квазифлексий, чтобы сравнение начиналось с самых длинных квазифлексий и кончая самыми короткими. Для исключения спорных случаев предусмотрен пользовательский интерфейс для «ручной» коррекции ссылки на список квазифлексий при помощи визуального сопоставления с образцами склонения и спряжения, по аналогии с методом, использованным в переводчике ПРОМТ[Светова].

В процессе сбора лексики системного аналитика словарь новой лексики формируется на основе иерарррхии расширенных диаграмм потоков данных (РДПД), при этом обработке подвергаются два свойства сущностей: имя и комментарий. При этом, кандидатами на помещение в словарь являются все слова, кроме уже имеющихся в словаре. Однако, в словарь помещаются только определенные словоформы, для которых успешно определен их тип словоизменения (выделена квазиоснова и

определена ссылка на список квазифлексий), или которые указал сам СА. Если СА пожелал занести в словарь слово, для которого не удается определить возможные формы или оно не имеет форм, то слово заносится в том виде, как встречается в тексте.

При определении морфологических характеристик слов в именах сущностей РДПД дополнительно учитываются следующие правила формирования имен сущностей ДПД[ЭЙТЭКС, 1993; Калянов, 1996]:

- имя внешней сущности чаще всего является существительным в единственном числе и именительном падежа;

- имя системы/подсистемы – обычно является простым предложением, состоящим из подлежащего в единственном числе и именительном падежа с определениями и дополнениями;

- имя процесса – обязательно предложение без подлежащего, состоящее из глагола в повелительном и неопределенной форме, за которым следует существительное в винительном падеже.

ПИЭ, который используется для формирования нового словаря, содержащего лексику эксперта, содержит «открытые вопросы», задаваемые экспертом самому себе, и ответы на вопросы, связанные друг с другом в последовательности, отражающей решение конкретной задачи, что обеспечивается гипертекстовой формой представления ПИЭ [Рыбина и др., 1997]. Поэтому ПИЭ можно рассматривать как частично семантически размеченный текст, что позволяет автоматизировать определение семантических категорий новых характеристик, строя гипотезы о принадлежности к семантическим категориям на основе категорий уже имеющихся в словаре слов из ответов эксперта на один и тот же вопрос и проверяя ее на других вопросах и ответах. При внесении в словарь данных о семантических категориях новой характеристики, полученной из ПИЭ, выполняется частичная корректировка модели управления предиката в вопросе или ответе эксперта для учета сочетаемости предиката с семантическими категориями новой характеристики.

Следует отметить, что для формирования новых словарей в комплексе АТ-ТЕХНОЛОГИЯ выполнялось на основе текстов, относящихся к проблемным областям медицинской и технической диагностики, проектирования в области машиностроения.

Рассмотрим следующий пример из области медицинской диагностики. В процессе интервьюирования эксперта был получен следующий вопрос с соответствующими ответами:

Вопрос: Какое нарушение состояния сосудов мозга?

Ответы:

спазм сосудов мозга

склероз сосудов мозга

тромбоз сосудов мозга

Все слова, кроме слова *тромбоз* уже занесены в словарь. Слова *спазм*, *склероз,* в соответствии с использованным классификатором, отнесены к категории «Заб» («Заболевание») , поэтому новое слово *тромбоз* при занесении в словарь также отнесено к категории «Заб». Для проверки правильности присвоения семантической категории новое слово ищется в остальных вопросах и ответах эксперта. Если оказывается, что слово *тромбоз* может быть отнесено также к другой семантической категории (например, «Осложнение заболевания»), то слово помечается как требующее дополнительного редактирования списка присвоенных ему семантических категорий.

Для решения задачи наполнения новых словарей в системе управления хранением и организации доступа к единому межведомственному банку данных «НЕВОД» (разработка фирмы ОПТИМА) удалось совместить семантический анализ текста с выявлением определенных статистических закономерностей.

При разработке системы, использованный ранее в комплексе АТ-ТЕХНОЛОГИЯ ЛП был адаптирован для выделения из текстовых документов, хранящихся в базе данных (БД), значений заданных атрибутов и занесения их в новые словари, использующиеся при поиске документов по уголовным делам в БД. В процессе адаптации ЛП, позволяющий ранее обрабатывать только отдельные простые предложения, был наделен способностью обрабатывать весь текст целиком, разбивая его на простые предложения и строя для каждого простого предложения предикатно-аргументную структуру.

Например, следующее предложение «Личность Иванова А.Б. удостоверена паспортом на его имя серии ХХ-МЮ № 123456, выданным 3 июля 1979 года 321 отделением милиции г. Москвы» преобразуется в следующие CAREL-выражения:

```
( удостоверить) ( МОД (глаг.сов.вид. крат. прич. ж.р. ед.ч.) )
    ( O ( личность ) (сущ. ж.р. ед.ч. В.п.) ( Пон Оду )
        ( H ( иванова ) (им.собств.)  )
        ( H ( а ) (им.собств.)  )
        ( H ( б ) (им.собств.)  )  )
    ( I ( паспорт ) (сущ. м.р. ед.ч. Т.п.) ( Пон Нео )
        ( H ( серия ) (сущ. ж.р. ед.ч. Р.п) ( Хар )
            ( H (хх-мю) (им.собств.) ) )
        ( H ( номер ) (сущ. м.р. ед.ч. И.п) ( Хар )
            ( H ( 123456 ) (число) ) ) )

( выдать) ( МОД (глаг.сов.вид. полное прич. м.р. ед.ч. Т.п.) )
    ( V ( 3 июля 1979 ) (число)  )
     ( L ( отделение ) (сущ. с.р. ед.ч. Т.п.) ( Пон Нео Мес)
        ( H ( 321 ) (число)  )
        ( P ( милиция ) (сущ. ж.р. ед.ч. Р.п) ( Хар Нео ) )
        ( H ( город ) (сущ. м.р. ед.ч. Р.п) ( Пон Нео Мес )
            ( H ( москва ) (сущ. ж.р. ед.ч. Р.п) ( Хар Нео Мес ) ) ) )
```

где, O - глубинный падеж «быть объектом действия»; I - глубинный падеж «быть инструментом действия»; L - глубинный падеж «быть местом действия»; H – обозначение связи типа «быть характеристикой»; P – обозначение связи типа «часть-целое»; МОД - зарезервированное слово для обозначения модальности; Пон, Хар, Оду, Нео, Мес - обозначения семантических категорий в принятой кодировке использованного в данном случае словаря.

В системе «НЕВОД» структура базовых словарей была сохранена, а структура новых словарей приспособлена для декларативного метода морфологического анализа [Андреев и др., 1998].

В процессе адаптации ЛП удалось автоматизировать процесс формирования модели управления предикатов, используя для этого массив примеров текстовых документов. При формировании моделей управления использовалось предположение о том, что предикаты, относящиеся к одной и той же семантической категории, имеют сходство моделей управления. На основе этого предположения был разработан алгоритм формирования моделей управления новых предикатов на основе уже имеющихся в словаре, состоящий из следующих этапов:

– на первом этапе, с помощью специального интерфейса пользователь заносит новый предикат в словарь и определяет для него семантические категории;

– на втором этапе, в словаре выполняется автоматический поиск предикатов, наиболее близких по своим семантическим категориям к новому предикату. Среди найденных предикатов выбирается модель управления, предполагающая наиболее широкую сочетаемость предиката, которая присваивается в качестве прототипа новому предикату;

– на третьем этапе, выполняется автоматическое сравнение и доопределение модели управления предиката на предлагаемых текстах.

Указанный алгоритм работает, если большинство слов, встречающихся в тексте, уже занесено словарь.

Например, пользователь заносит в словарь предикат *возобновить, который может* встречаться в таких предложениях, как «Предварительное следствие по уголовному делу № 12345 возобновить, приняв его к своему производству» или «31 сентября 1999 возобновлено предварительное следствие». В процессе занесения в словарь пользователь определяет одну или несколько семантических категорий, к которым относится предикат *возобновить.* Фрагмент описания одного семантического класса "Действие, процесс" имеет следующий вид:

```
{
    Дей:"Действие, процесс"
    {
        ........
        Нез:"Незаконное действие",
        Дел:"Действие над уголовным делом",
        Про:"Действие в соответствии с уголовным делом",
        ........
    }
}
```

При занесении предиката *возобновить* в словарь пользователь определяет, что он относится к категории "Действие над уголовным делом", обозначенной как «Дел». При поиске в словаре предикатов найдены предикаты, также относящиеся к «Дел»: *возбудить*, *прекратить* и др. После сравнения моделей управления найденных предикатов в качестве прототипа предикату *возобновить* поставлена в соответствие модель управления предиката *возбудить*, так как в отличие от других найденных предикатов предикат *возбудить* встречается в сочетании с фактом возбуждения уголовного дела, которому присваивается глубинный падеж Т - «быть темой действия» (например, «Настоящее уголовное дело было возбуждено 01 сентября 2000 года РУВД ЦАО г. Москвы по факту незаконной банковской деятельности»). Дальнейшее сравнение прототипа модели управления предиката *возобновить* с имеющимися тестовыми документами показывает, что наличие глубинного падежа Т нехарактерно для данного предиката, поэтому для формирования окончательной модели управления фильтр для Т удаляется из прототипа модели управления.

Для усиления возможностей выявления новой лексики в системе «НЕВОД» использовались алгоритмы выявления статистических закономерностей встречаемости слов. Примером может служить описанная ниже статистическая закономерность появления в текстах фамилий обвиняемых.

Исследования лексики текстов уголовных дел показали, что она может быть достаточно хорошо классифицирована по частоте использования слов. Их можно разделить на следующие группы:

1) слова, постоянно встречающиеся во всех текстах;
2) слова, часто встречающиеся в отдельно взятом тексте, но редко встречающиеся во всех текстах;
3) слова, редко встречающиеся во всех тестах.

К первой группе относится, как правило, специальная лексика (грантпост, обвиняемый, правительство, следствие и т.д.), собственные имена (Москва, Россия, Федерация и.т.д.), сокращения (УК, УПК, РФ, и т.д.) и некоторые предлоги, союзы и частицы. Ко второй группе относятся, чаще всего, фамилии обвиняемых. К третьей группе относится остальная, чаще всего, общезначимая лексика. Экспериментально установлено значение порога встречаемости слова во всех или в отдельно взятом тексте, приблизительно равное 0.2% от общего количества слов текста, которое использовано, чтобы отнести слова к одной из трех групп. Для выделения фамилии обвиняемого слова из первой группы предварительно заносятся в новый словарь и используются при обработке текста для исключения из списка слов, частота появления которых в тексте больше экспериментально установленного порога.

Литература

[Автоматизация анализа научного текста, 1984] Автоматизация анализа научного текста. Киев,"Наукова думка",1984.

[Андреев и др., 1998] Андреев А.М., Березкин Д.В, Брик А.В. Лингвистический процессор для информационно-поисковой системы. Компьютерная хроника , 1998, № 11, С. 79-100.

[Калянов, 1996] Калянов Г.Н. Структурный системный анализ (автоматизация и применение). М., «Лори», 1996.

[Колобашкина и др., 1996] Колобашкина М.В. Рыбина Г.В. Сергиевская О.Г. Смирнов В.В. Задачно-ориентированная методология приобретения знаний для компьютерного построения интегрированных экспертных систем // В кн.: КИИ-96 Пятая нац. конференция с межд. участием «Искусственный интеллект-96». Сборник научных трудов в трех томах. Том 2. Казань, 1996. С. 270-274.

[Кустикова и др., 1998] Кустикова И.А., Рыбина Г.В., Смирнов В.В. Об одном подходе к автоматизированному построению базы знаний для интегрированных экспертных систем: аспекты тестирования. В кн.: КИИ’98 Шестая нац. конференция по искусственному интеллекту с межд. участием. Сбоник научных трудов в трех томах. Том 2. Пущино, 1998. С. 138-144.

[Пышагин и др., 1996] Пышагин. С.В., Рыбина Г.В., Смирнов В.В. Инструментальный комплекс АТ-ТЕХНОЛОГИЯ для поддержки проектирования интегрированных экспертных систем // В кн.: КИИ-96 Пятая нац. конференция с межд. участием «Искусственный интеллект-96». Сборник научных трудов в трех томах. Том 3. Казань, 1996. С. 522-527.

[Рыбина, 1990] Рыбина Г.В. Модель диалога в естественно-языковой системе ДИСАР. В кн.: Автоматизированная информационная технология. М.: Энергоатомиздат, 1990, С.29-36.

[Рыбина, 1997] Рыбина Г.В. Задачно-ориентированная методология автоматизированного построения интегрированных экспертных систем для статических проблемных областей // Изв. РАН. Теория и системы управления. № 5, 1997. С129-137.

[Рыбина и др., 1997] Рыбина Г.В., Пышагин С.В., Смирнов В.В., Чабаев А.В., Пашина И.А. Автоматизированное построение интегрированных экспертных систем на основе средств

инструментального комплекса АТ-ТЕХНОЛОГИЯ (версия MS-Windows). В кн.: Международная летняя школа-семинар по искусственному интеллекту для студентов, аспирантов и молодых ученых. Сборник трудов. Минск , 1997, С. 119-137.

[Рыбина и др., 1998] Рыбина Г.В., Смирнов В.В., Кустикова И.А., Ледовская Т.В., Солонович Е.А., Файбисович М.А., Гриценко Ф.А. Автоматизированное построение базы знаний для интегрированных экспертных систем на основе средств комплекса АТ-ТЕХНОЛОГИЯ // В кн.: Научная сессия МИФИ–98. Сборник научных трудов. Ч.5. М:МИФИ, 1998, С. 34-37.

[Рыбина и др., 1999] Рыбина Г.В., Смирнов В.В., Левин Д.Е. Автоматизированное извлечение знаний, содержащих НЕ-факторы. В кн.: Научная сессия МИФИ–99. Сборник научных трудов. Том 7. М:МИФИ, 1999, С. 184-185.

[Рыбина и др., 2001] Рыбина Г.В., Душкин Р.В., Душкина Е..Н. О новых возможностях лингвистического процессора инструментального комплекса АТ-ТЕХНОЛОГИЯ // В кн.: Научная сессия МИФИ–2001. Сборник научных трудов. Том 3. М:МИФИ, 2001, С. 134-135.

[Рыбина, Душкин и др., 2001] Рыбина Г.В. Душкин Р.В. Душкина Е.Н. Лингвистические аспекты извлечения знаний, содержащих неопределенность, неточность и нечеткость // В кн.: Международный научно-практический семинар «Интегрированные модели и мягкие вычисления в искусственном интеллекте». Сборник трудов. М.: Физматлит: 2001. С. 168-172.

[Рыбина, Пышагин и др., 2001] Рыбина Г.В., Пышагин С.В., Смирнов В.В., Левин Д.Е., Душкин, Р.В. Некоторые особенности Windows-версии инструментального комплекса АТ-ТЕХНОЛОГИЯ. В кн.: Научная сессия МИФИ–2001. Сборник научных трудов. Том 10. Телекоммуникации и новые информационные технологии в образовании. М:МИФИ, 2001, С. 58-59.

[Светова] Светова С. Опыт создания средств редактирования словаря пользователями системы машинного перевода семейства ПРОМТ. (КОМПАНИЯ "ПРОМТ", С.-Петербург).

URL: http://www.promt.ru/mtw/reports/dialogue99.phtml.

[ЭЙТЭКС, 1993] CASE.Аналитик. Версия 1.1. Руководство аналитика. Москва, научно-техническое предприятие ЭЙТЭКС, 1993.

THE REUSE OF DICTIONARIES IN INFILLING NEW ONES

Vitaly V. Smirnov
Moscow State Engineering Physics Institute (Tech. University),
115409 Moscow, Kashirskoe Shosse, 31, MEPhI,
Vitaly_Smirnov@mail.ru

This report observes the experience of automation of lexicon extraction from texts to infilling new dictionaries, using a base lexicon dictionary, statistics of words appearance and compatibility. New dictionaries are formed and used by specific linguistic processor (LP), the main function of which is natural language (NL) phrase transmission to the predicate-argument structures in CAREL language. LP is used in AT-TECHNOLOGY workbench to solve a number of problems connected with NL processing and in NEVOD system to given attribute values detection in criminal case documents and to find documents in data base.

New words morphological characteristics are determined by means of comparison with the table of quasi-inflection. At the beginning of the comparison the longest quasi-inflection are used and the shortest ones are used in the end which is provided by indexing of quasi-inflection.

One of the problems solved by means of LP in AT-TECHNOLOGY complex is the new lexicon extraction from expert answers, received by interview. While the lexicon extraction semantic categories of new words are determined by comparison with semantic categories of the words in the dictionary and predicate control models are corrected. To supplement LP dictionaries in NEVOD system by new predicates forming control models algorithm using the semantic classifier is developed.