

УДК 681.3.01

И.П. Кузнецов, Н.В. Сомин, Е.Б. Козеренко, А.Г. Мацкевич

Институт проблем информатики РАН, г. Москва, Россия
igor-kuz@mtu-net.ru, somin@post.ru, kozerenko@mail.ru

Особенности лексико-морфологического анализа в задачах извлечения структур знаний из текстов естественного языка

Рассматривается класс объектно-ориентированных лингвистических процессоров, выделяющих структуры знаний из текстов естественного языка (ЕЯ). Важной компонентой таких систем является блок лексико-морфологического анализа. В процессе разработки приложений этот блок постоянно совершенствовался и приобрел много новых функций, выходящих за рамки возможностей существующих блоков подобного типа. Данный блок генерирует лексические, морфологические, семантические признаки слов, выявляет простейшие формы естественного языка, имеет специальные средства настройки на предметную область и на особенности текстов ЕЯ. В работе рассматриваются эти функции.

Введение

На протяжении многих лет в ИПИ РАН активно развивается область, связанная с построением систем, обеспечивающих извлечения полезной информации из текстов естественного языка (ЕЯ) с формированием структур знаний и их использованием для решения прикладных задач – поисковых, логико-аналитических. Для таких систем требовались специальные языки представления знаний и инструментальные средства их обработки. Учитывался тот факт, что язык – это структурный объект на всех его уровнях: от поверхностного до семантического. Для обработки конструкций языка были созданы язык расширенных семантических сетей (РСС), обеспечивающий представление текстов ЕЯ на уровне структур знаний с любой требуемой точностью, и язык ДЕКЛ – для преобразования структур в виде РСС [1-3].

Важной составляющей логико-аналитических систем, имеющих дело с информацией на ЕЯ, является лингвистический процессор, отображающий тексты ЕЯ на структуры знаний [4-6]. При разработке таких процессоров учитывался тот факт, что определенные категории пользователей интересуются конкретной информацией, которая встречается в текстах ЕЯ. Нужно извлекать из текстов только эту информацию. Данное направление возникло в связи с прикладными разработками для ГУВД г. Москвы. Их проблемы заключались в наличии потоков документов на ЕЯ (сводок происшествий, справок по уголовным делам, обвинительных заключений и др.), в которых было много полезной информации. Это фигуранты, их адреса, телефоны, оружие, автотранспорт и др. Будем называть их **информационными объектами** (другое название – сущности). Следователей и аналитиков интересовали именно такого сорта объекты и связи между ними. Использование типовых БД требовало громадной работы для их заполнения.

В связи с этим в ИПИ РАН была инициирована работа по созданию лингвистических процессоров (ЛП), обеспечивающих автоматическое выделение их текстов ЕЯ информационных объектов и связей с формированием структур знаний. Такие ЛП

были названы **объектно-ориентированными**. Были созданы системы «Криминал», «Аналитик» и др., обеспечивающие автоматическое извлечение структур знаний из текстов ЕЯ и их использование для решения логико-аналитических задач [3], [6], [7]. Важной компонентой ЛП является блок **лексико-морфологического анализа** (ЛМА), который анализирует текст и строит семантическую сеть (РСС), названную **пространственной структурой текста** (ПС-текста) [6]. Последняя обрабатывается блоком **синтактико-семантического анализа** (ССА), который (на языке ДЕКЛ) анализирует ПС-текста и формирует на РСС структуру, представляющую объекты и связи между ними. Такие структуры образуют базу знаний (БЗ).

Отметим, что блок ЛМА написан на языке Си++, при использовании которого на определенных этапах формализации текстов возникают существенные трудности. В то же время чем больше функций берет на себя блок ЛМА, тем в большей степени снимает трудности дальнейшего процесса формализации, который осуществляется блоком ССА [3], [5], [6].

1 Компоненты объектно-ориентированных лингвистических процессоров

Опыт многих разработчиков показывает, что при автоматическом анализе потока документов учесть все формы и особенности ЕЯ и построить сколь либо полную «модель языка» – неразрешимая задача. Поэтому требуется постоянное совершенствование ЛП. В связи с этим перспективным представляется направление, когда программа объектно-ориентированного ЛП отделяется от **лингвистических знаний** (ЛЗ). Последние определяют всю процедуру анализа (см. ниже). ЛЗ имеют вид декларативных структур, которые легко менять и настраивать. В нашем случае роль таких структур выполняют фрагменты РСС [3], [5], [6]. Настройка ЛП осуществляется только за счет разработки ЛЗ.

Задача ЛП – поддерживать ЛЗ. При использовании подобных ЛП облегчается настройка на корпуса текстов, особенности предметной области. Корректировать ЛЗ может человек, обученный формализму РСС и знакомый с элементами математической лингвистики. Ему не нужно уметь программировать.

Рассмотрим основные компоненты объектно-ориентированных ЛП.

1.1. Блок лексико-морфологического анализа (ЛМА) выделяет из документа слова и предложения и выдает в виде семантической сети (ПС-документа), представляющей последовательность компонент (слов в нормальной форме, чисел, знаков) и их основные признаки. Блок ЛМА имеет три основных подсистемы:

– Лексический анализатор, который ответственен за правильное деление входного текстового потока на абзацы, предложения и слова (формирует лексические признаки слов);

– Морфологический анализатор, осуществляющий морфологический анализ всех слов текста (приводит слова в нормальную форму и формирует для них морфологические признаки).

Блок ЛМА имеет свои лингвистические знания (ЛЗ) – средства **параметрической настройки**, позволяющие учитывать разнообразие текстовой типологии, и набор **предметных словарей** (словарь стран, регионов России, имен, профессий и др.) для придания словам и словосочетаниям дополнительных семантических признаков [4], [5].

1.2. Блок синтактико-семантического анализа (ССА) путем анализа ПС-документа выделяет объекты и связи. На их основе строит другую семантическую сеть, представляющую **семантическую структуру документа** (СС-документа), называемую

также **содержательным портретом** [3], [6], [7]. Этот блок включает в себя **базу лингвистических знаний** (ЛЗ), которая содержит правила анализа текста во внутреннем представлении (РСС). Они определяют работу ЛП.

Блок ССА управляется ЛЗ, за счёт которых обеспечивается:

- извлечение информационных объектов (лиц, организаций, событий, их места);
- выявление связей объектов; например, связей лиц с организациями, адресами и др.;
- анализ глагольных форм, причастных и деепричастных оборотов с выявлением фактов участия объектов в тех или иных действиях;
- идентификация объектов с учетом анафорических ссылок и сокращенных наименований;
- выявление связей действий с их местом или временем (где и когда происходило данное действие или событие);
- анализ причинно-следственных и временных связей между действиями и событиями.

Особенности блока ССА описаны во многих статьях [3], [5], [6]. Гораздо меньше внимания уделялось описанию работы блока ЛМА. В данной статье будет восполнен этот пробел.

Блок ЛМА [4], [5] основан на традиционной для таких блоков схеме словарей. Однако, помимо этого, в блоке ЛМА присутствует еще **словарь обобщенных основ**, позволяющий обрабатывать и новые слова (п. 4).

Блок ЛМА приводит слова в нормальную форму и присваивает им признаки, которые делятся на три группы:

- лексические признаки (слово с большой буквы, большими буквами, с точкой на конце или это отдельная буква и др.)
- морфологические признаки (грамматическая категория слова, число для существительных и т.д.);
- семантические признаки (имя, организация, оружие и др., а также ключевые слова, относящиеся к соответствующему типу объектов).

Предусмотренный лексикографический анализ обеспечивает автоматическое деление текста на самостоятельные части (например, выделение документов из сводок) и определение начала и конца предложения, а также начала и конца абзаца.

Выходная информация блока ЛМА (т.е. ПС-текста) сохраняет порядок предложений в тексте, разделяя их фрагментами типа SENT, и порядок слов в предложении. При этом каждое слово представляется с его признаками (п. 6).

2 Прикладные области и тексты

В настоящее время имеется большой опыт использования объектно-ориентированных ЛП в прикладных областях, где требуется выделение различных объектов из корпусов текстов со своими особенностями. В данном разделе мы постараемся обобщить эти особенности и связанные с ними трудности, которые требовали постоянного совершенствования блока ЛМА. Мы имели дело с такими предметными областями и текстами:

2.1. Документы криминальной милиции. Работа делалась по заказу ГУВД г. Москвы [3], [7]. Была создана система «Криминал», в БЗ которой были введены: сводки происшествий (более 500 тыс. происшествий), справки по уголовным делам, обвинительные заключения, записные книжки фигурантов и др. Система обеспечивает

выделение фигурантов, их примет, связей, организаций, дат, документов, номеров счетов, оружия (всего до 40 типов объектов) с указанием характера их участия в криминальных действиях.

2.2. Резюме (для приема на работу) на русском и английском языках. Работа имела целью автоматическую обработку архивов произвольно написанных резюме и их представление в формате сайта одной из компаний, осуществляющей поиск работы для клиентов [3]. Была создана система, выделяющая из резюме атрибуты человека, места его работы, учебы, соответствующие периоды времени, знание языков и т.д. Система отлаживалась на выборках в различных областях: информационные технологии, банковское дело, финансы, юриспруденция и др. Система работала на сайте упомянутой компании, чтобы автоматически переводить резюме пользователей, поступающих через Интернет, в формат сайта.

2.3. Документы о терроризме на русском языке. Работа носила инициативный характер с целью внедрения в крупный проект. Система дополнительно выделяла руководящих лиц, правительственные организации, террористов (как свойство фигурантов), террористические организации, орудия преступления, время и место событий и т.д., а также связи и участие лиц в тех или иных действиях.

2.4. Документы о памятниках культуры. Работа делалась для Министерства культуры. Система выделяет из текстов тип памятника (скульптура, монумент), кто является автором, создателем, время, место и многое другое.

Во всех случаях (за счет средств настройки блоков ЛМА и ССА) удавалось добиться требуемого качества работы ЛП [3], [6], [7].

Отметим высокое разнообразие перечисленных предметных областей, которое определяется не только различием выделяемых объектов и связей. Еще большие отличия можно наблюдать в «стиле» текстовых сообщений, связанных с предметными областями. В понятие «стиль» мы включаем весь комплекс особенностей, присущих определенной группе текстов. Сюда входят:

- лексика предметной области, включая всю совокупность специфических терминов предметной области;
- коммуникативный тип текста: художественное произведение, техническая или аналитическая статья, новостное сообщение, приказ, PR-текст (например реклама);
- структурный тип текста: связный текст, список, таблица, математическая формула;
- инструмент создания текста (имеется в виду текстовый редактор или генератор текста, с помощью которого получен текст);
- способ грамматического оформления текста, под которым понимается следование стандартным правилам орфографии языка (проставление необходимых знаков препинания и разделителей, позволяющих структурировать текст);
- следование принятой в языке орфографии, что выражается в количестве орфографических ошибок или нарочитом введении искаженной лексики.

Отметим резкое увеличение разнообразия текстовой типологии, с которой мы столкнулись в различных предметных областях. В значительной степени это вызвано бурным распространением Интернета и тем фактом, что порождение текстов все в большей мере стали осуществлять люди различной степени подготовки и грамотности. Как следствие – наличие значительного количества специальных разделителей, отсутствие знаков препинания, большое количество сокращений, ошибок и многое другое. Отсюда следуют дополнительные требования к компонентам блока ЛМА и средствам их настройки. Рассмотрим их подробнее.

3 Особенности лексического анализатора

Лексический анализатор имеет дело с целым рядом взаимосвязанных задач, решение которых совершенно необходимо для успешной работы всего ЛПА. Рассмотрим их особенности.

Прежде всего, решается **задача структуризации текста**. Дело в том, что текст в современной информационной среде – сложно структурированный объект. И его структура должна быть распознана и аккуратно передана блоку ССА. От правильного распознавания структуры текста в значительной степени зависит корректность всего анализа по извлечению знаний. Поэтому задача структуризации распадается на цепочку локальных задач.

3.1. Трудности выделения лексем. Рассмотрим трудности выделения из входного потока лексем: слов, знаков препинания, разного рода разделителей и др. Современный деловой текст содержит большое количество лексем, являющихся техническими, административными и фирменными названиями, телефонами, шифрами, номерами автомобилей, адресами электронной почты и Интернета, содержащими цифры, буквы и разделители практически в произвольной комбинации. Такие знаки, как «-», «.» и «,», доставляют много хлопот при их анализе, в одних случаях являясь разделителями лексем, а в других – нет.

3.2. Задача выделения предложений. Ввиду огромного разнообразия текстовых «стилей», по отношению к современным текстам становится трудно говорить о предложении. Скорее следует говорить о «сильносвязанных» отрезках текста, в которых идет речь об одном объекте или одной ситуации, в которой участвуют несколько взаимодействующих объектов. В результате само понятие «предложение» резко расширяется, включая в себя, помимо обычных предложений (с точкой в конце), еще массу различных текстовых отрывков: ячеек таблицы, элементов списка и прочих, грамматическое оформление которых нетрадиционно.

3.3. Задача выделения абзацев. Абзацем мы называем отрезок текста из одного или нескольких предложений, связанных единой темой. Расплывчатость этого определения позволяет трактовать его достаточно широко. Однако для блока ССА понятие абзаца является весьма важным, поскольку многие его механизмы направлены именно на идентификацию и совмещение объектов внутри одной темы. Лексический анализатор содержит в своем составе ряд алгоритмов, выделяющих абзацы, причем – разных типов.

Как оказалось, задачи выделения предложений и абзацев весьма нетривиальны. Трудности выделения абзацев главным образом связаны с тем, что хорошо различимые разделители абзаца – пустые строки, отступы, границы клеток таблицы – теряются или искажаются при преобразовании текстов. Но гораздо большие трудности возникают при идентификации предложений. Дело в том, что современные пользователи Интернета вообще не считают необходимым ставить точки в конце предложения. В то же время точка активно используется в качестве ограничителя сокращений, разделителя между частями электронного адреса, многозначного числа, банковского номера и др. Кроме того, разделителем предложения может являться не только точка, но и другие знаки («;», «:», «!», «?», «|» и т.д.). В результате задача разбиения текста на предложения становится просто головоломной шарадой, требующей учета массы разного рода частных правил и исключений.

3.4. Проблемы унификации текста. Естественный язык – система необычайно многовариантная. Задача лексического анализатора: унифицировать написание отдель-

ных слов и сокращений, привести к стандартной форме написания ряда стандартных словосочетаний. Трудности возникают при выявлении наиболее употребительных лексем и словосочетаний, требующих унификации.

К этим трудностям добавляется проблема обнаружения и (по возможности) исправления **опечаток и грамматических ошибок**. В современных текстах их громадное количество, и бороться с ними – задача из сложнейших. Кроме того, в современных текстах, особенно из Интернета, намечается тенденция нарочитого переделывания и перевирания слов, типа «*ацкий ужос*» или «*падстол*». Начинает формироваться целая интернетная «феня». В связи с этим потребуются постоянная корректировка языковых словарей и правил составления предложений.

Еще одна важная функция лексического анализатора – определение **лексических признаков** слов. Примеры такого рода признаков: «слово из кириллицы с прописной буквы», «слово из кириллицы из прописных букв», «разделитель», «слово из латинских букв» и проч., всего – около 20 лексических признаков. Лексические типы являются важной дополнительной информацией, облегчающей работу как морфологического анализатора, так и блока ССА.

Наконец, лексический анализатор для ряда слов способен выполнить семантический анализ, определяя по формальному виду слова его **семантическую категорию**. К этому случаю относятся сокращения имен и отчеств: прописная буква, за которой идет «.». Например «*А.*», «*Н.*», «*Ж.*». Еще примеры идентифицируемых семантических классов: «адрес электронной почты», «Интернет-адрес» (URL), «целое число», «число с дробной частью». Собственно, определение семантического класса каждого слова или словосочетания является одной из задач всего ЛП. И чем раньше такой класс будет определен, тем легче дальнейший анализ.

4 Особенности морфологического анализатора

Задача морфологического анализатора – нормализация слов, определение морфологических признаков лексем, а также (в ряде случаев) нахождение их семантических классов. Отметим, что к настоящему времени разработан целый ряд морфологических анализаторов русского языка [8], [9].

4.1. Схема анализа. Первоначально была реализована базовая схема анализа [6]. Считается, что каждое слово имеет постоянную часть (основу) и переменную часть. Последняя образует словоизменительную парадигму или класс окончаний. Были накоплены два словаря: словарь классов окончаний (СКО), в котором хранятся все возможные парадигмы русского языка и словарь основ (СО), в котором хранятся основы слов со ссылками на соответствующий класс окончаний.

Например, слово «*бытие*» имеет основу «*быти*» и класс окончаний за номером 1759, содержащий окончания в именительном, родительном, дательном, винительном, творительном и предложном падежах, а именно: «*е*», «*я*», «*ю*», «*е*», «*ем*», «*и*» (множественного числа это слово не имеет). Соответственно в СО имеется запись «*быти 1759*», а в СКО под номером 1759 закодирована парадигма с указанными окончаниями.

Отметим, что в общем случае в СО может быть несколько записей с одинаковой основой (но с разными классами окончаний), а на один и тот же класс окончаний может ссылаться несколько слов с разными основами. Возможны случаи пустой основы (пример: «*хорошо*»-«*лучше*») и пустого класса окончаний (для неизменяемых слов). Кроме основы и вариантов окончаний, в СКО хранятся морфологические признаки,

соответствующие определенному классу окончаний в целом (постоянная морфологическая информация) и каждому окончанию парадигмы в отдельности (переменная морфологическая информация). Так, для класса 1759 в качестве постоянной информации хранятся признаки существительного, среднего рода, неодушевленности и второго склонения, а для каждого окончания хранится признак соответствующего падежа.

Алгоритм морфологического анализа при наличии данных словарей сводится к следующему. Для слова рассматриваются все варианты его разбиения на основу и окончание. Если для данного варианта разбиения находится основа, а в соответствующем ей классе окончаний находится вариант окончания, то данный морфологический разбор является корректным и слово получает морфологические признаки, взятые из постоянной и переменной частей морфологической информации. В общем случае может быть найдено и выдано несколько вариантов морфологического разбора, что известно как морфологическая омонимия.

4.2. Морфологический анализ незнакомых слов. В принципе предложенная схема анализа вполне корректна. Однако на практике ее успешное использование достаточно проблематично. Дело в том, что такая схема предполагает ручную разработку обоих словарей. И заметим – не только первоначальную разработку, но и их постоянное пополнение. Последнее обстоятельство особенно неприятно: в русском языке – более 100 тыс. слов общеупотребительного назначения и миллионы специальных терминов. Кроме того, за последнее время в русскоязычных текстах стало использоваться огромное количество англоязычных слов, которые никогда не входили в классические словари русского языка. Фактически требовалось ежедневное пополнение словаря.

Выход из описанной ситуации известен – обработка незнакомых системе слов «по аналогии» [8], [9]. В нашей реализации этого метода использовался третий словарь – «*словарь хвостов основ*» (СХО). В словарь записываются все 1-буквенные, 2-буквенные, 3-буквенные и т.д. «хвосты» основ (первые буквы основ отбрасываются) с указанием соответствующего класса окончаний. Было решено, что в СХО не будет одинаковых «хвостов», а его класс окончаний вычисляется из статистических соотношений – по максимуму основ в СО, имеющих данный «хвост» и данный класс окончаний. Если слово не находится в словаре СО, то та же схема анализа повторяется, но уже с помощью пары словарей СХО-СКО.

В реализации словари СО и СХО были слиты в один словарь, за которым закрепилось название обобщенного словаря основ (ОСО), в результате чего все варианты анализа, – как точные, так и по аналогии, – выявляются за один проход по словарю.

4.3. Способы устранения морфологической омонимии. Ясно, что использование обобщенного словаря основ ОСО может приводить к лишним вариантам морфологического анализа. Было предложено два достаточно эффективных способа борьбы с морфологической омонимией.

Первый способ – эмпирический алгоритм, отбрасывающий наименее вероятные варианты морфологического анализа. Такая «зачистка» вариантов выполняется по многим критериям, учитывающим наличие слова в СО, длину основы с СХО, часть речи. Кроме того, эмпирический алгоритм расставляет все варианты разбора в порядке их вероятности.

Второй способ – частичный синтаксический анализ, позволяющий отбросить варианты морфологического анализа, которые не удовлетворяют критериям согласования слов. Для этого было реализовано распознавание двух конструкций: полного согласования и генетической цепочки [4].

5 Предметные словари

Предметные словари (стран, имен собственных, организаций, профессий, видов оружия и др.) состоят из терминов. Множество словарей образует систему.

Система предметных словарей (СПС) предназначена для распознавания в тексте слов и словосочетаний, специфичных для конкретной предметной области. Им присваиваются признаки принадлежности к определенной семантической категории. Будем называть этот процесс идентификацией терминов словаря. Такая принадлежность является основой выделения объекта. В предметном словаре может быть или термин, представляющий объект определенного типа (таких объектов может быть достаточно много), или характеристическое слово, опираясь на которое можно начинать распознавание объекта – на уровне синтактико-семантического анализа.

Как показывает опыт, СПС является необходимой компонентой любого объектно-ориентированного ЛП. В нашей разработке СПС встроена в блок ЛМА. Причина этого – главным образом в быстрой работе. Поиск в СПС предполагает частые обращения к ней, а потому требуется высокая эффективность поиска, чего трудно достичь без использования универсальных языков программирования. В нашем случае программное обеспечение СПС написано на Си++.

Структурно СПС состоит из произвольного количества **словарей**, представляющих собой определенный семантический класс. В каждом из словарей может содержаться произвольное количество **словарных записей**. Под записью в тривиальном случае понимается термин (однословный или многословный). Однако простыми терминами словарные объекты не ограничиваются. Допускаются записи в виде **словарных шаблонов**, описывающих группу терминов (п. 5.2). В настоящее время разработаны более 20 предметных словарей; среди них: «улицы г. Москвы», «террористические организации», «оружие», «известные личности» и т.д.

5.1. Требования к предметным словарям. К СПС, помимо эффективности, предъявляются еще ряд требований, важнейшим из которых является **требование вариативности поиска**. Должна быть предусмотрена корректная обработка случаев, когда написание термина в тексте так или иначе не соответствует его каноническому виду в словаре. Основная трудность – когда имеет место множество вариантов употребления одного и того же термина. Их нужно приводить к одному виду. Рассмотрим примеры.

Как правило, названия улиц записаны в именительном падеже. Например, «*проживает по адресу Б. Академическая ул. д. 6-18*». Иногда встречается дательный падеж: «*по Б. Академической*». Гораздо более усложняет дело вариативность сокращений и перестановки слов. Например, канонический вид названия одной из улиц Москвы – «*Щипковский 1-й пер.*». Однако встречаются в текстах написания: «*1-й Щипковский пер.*», «*1-ый Щипковский переулок*», «*п-к 1-ый Щипковский*» и другие варианты. Отметим, что возможна не только перестановка и вариативное написание слов, но и выпадение или добавление слов. Например, «*Туполева Академика наб.*» может быть названа как «*набережная Туполева*», а в название «*Тихий туп.*» иногда добавляют пояснение «*ул. Тихий туп.*». Кроме того, некоторые сокращения, применяемые авторами текстов, далеко не однозначны. Например «*С.*» может означать «*Северный*» или «*Старый*»; «*Б.*» может означать «*Большой*», а может быть сокращением имени, например «*ул. Б. Галушкина*».

5.2. Возможности предметных словарей. Подключение новых словарей может значительно усилить ЛП в плане выделения объектов. Однако для того, чтобы словари

в самом деле стали действенным и удобным механизмом, необходимо, чтобы они обладали рядом нетривиальных возможностей.

В нашей версии СПС реализованы несколько таких возможностей.

Во-первых, идентификация термина в любом числе и падеже. Например, если в словаре есть термин «*программный продукт*», то в тексте будут распознаваться и соответствующим образом идентифицироваться термины «*программного продукта*», «*программных продуктов*» и т.д. Распознавание выполняет программное обеспечение системы предметных словарей, использующее блок морфологического анализа.

Во-вторых, допускается несколько вариантов написания одного и того же термина. Дело в том, что в средствах СМИ и многих других текстах пользуются различными вариантами именования одного и того же объекта, в том числе сокращенным описанием. Например, если в тексте встретилось *Путин, Меркель, президент Франции* и т.д., то понятно, о ком идет речь. Для приведения таких словосочетаний к стандартному виду в словари введена специальная запись. Например, в словаре ФИО может иметь место запись:

Меркель Ангела
= *Ангела Меркель*
= *А. Меркель*
= *Меркель*

В данном примере основной термин – «*Меркель Ангела*». К нему будут приводиться все остальные написания этого имени, записанные после символа «*=*». Эта возможность особенно эффективна при выявлении не только ФИО известных деятелей, но и названий организаций (включая их сокращения), географических названий и др. При этом блок ССА осуществляет дополнительную фильтрацию, например, когда в тексте несколько лиц с фамилией *Меркель* или рядом со словом *Меркель* стоит какое-либо имя, не представленное в предметном словаре.

В-третьих, в предметные словари введена возможность описания группы терминов, у которых лишь первое слово фиксировано, а остальные могут быть описаны с помощью совокупности признаков (лексических и морфологических). Реализованы так называемые **словарные шаблоны**. Например, в словаре допустима запись:

заведующий {NOUN, KEM}.

Такая запись в словаре профессий означает, что подходящими под этот шаблон терминами могут быть все словосочетания, начинающиеся со слова «*заведующий*», за которым идет существительное (NOUN) в творительном падеже (KEM): «*заведующий складом*», «*заведующий библиотеками*» и т.д. Кроме того, в качестве шаблона можно употреблять имя другого (или того же самого) словаря. Фактически на словари возлагаются элементы синтаксического анализа, позволяющие значительно уменьшить количество записей в словаре, а также облегчить работу блока ССА.

В-четвертых, имеется возможность управлять лексическим и морфологическим анализами в процессе распознавания терминов словарей. Так, например, в словаре террористических организаций может быть указано:

Организация эта\
= *ЭТА!*

Это означает, что, благодаря признаку «\», слово «*эта*» в процессе идентификации морфологическому анализу не подвергается (т.е. его каноническая форма совпадает с написанием). И кроме того, благодаря признаку «!», идентификация совершается, если в тексте слово «*ЭТА*» записано прописными буквами. Эти возможности позволяют повысить точность распознавания, отсеивая ложные вхождения.

Отметим, что язык записи терминов в словарях чрезвычайно прост. Термин пишется в своей канонической форме на отдельной строке (включая, разумеется, указанные выше дополнительные возможности). Поэтому ввод новых терминов или даже создание новых словарей может быть выполнено пользователем или оператором-лингвистом, не знакомым с особенностями работы ЛП.

Помимо указанных возможностей имеется еще ряд специальных операторов настройки, позволяющих управлять идентификацией терминов для тех или иных словарей.

6 Представление пространственных структур

Текст ЕЯ – это сложный структурный объект, который в процессе его формализации проходит множество уровней преобразования. На первом уровне работает блок ЛМА, который формирует РСС, называемую **пространственной структурой текста** (ПС-текста). Далее следуют преобразования, осуществляемые блоком ССА, которые приводят к формированию **семантической структуры** (СС-текста) для БЗ.

Рассмотрим особенности ПС-текста. Информация об абзацах и предложениях представляется в виде фрагмента SENT, с помощью которого представляется:

- позиция первого слова предложения относительно начала входного потока;
- признак начала абзаца и количество разделительных строк;
- номер строки, на которой расположено первое слово предложения.

Для каждого слова (и для каждого варианта его разбора) блок выдает фрагменты типа LR, задающие последовательность слов. В каждом из фрагментов представлено: нормализованное слово и его порядковый номер. Далее следуют его признаки. Вот некоторые из них: NAME0 – слово начинается с прописной буквы, HEAD_ – слово полностью состоит из прописных букв, NAME1 – инициалы, POINT – пункт, HEAD_1 – слово с прописной буквой, NUM) – целое число, NUM_F – число с дробной частью, ENGL – слово из букв латинского алфавита, WEB_C – URL (адрес Интернет), MAIL_E – адрес электронной почты, FIRST_ – признак первого слова на новой строке, LETT – слово из одной буквы и т.д. (морфологические и семантические признаки).

Фрагменты типа LR и SENT вместе с выделенными признаками – это семантическая сеть (РСС), которая в дальнейшем проходит множество уровней преобразования, осуществляемое блоком ССА.

В общем случае блок ЛМА выдает несколько вариантов разбора. Например, слово «стекло» является и существительным, и глаголом. Тогда в ПС-текста, помимо фрагмента LR для первого варианта разбора, генерируются фрагменты LD (с их признаками) для других вариантов. Отсев вариантов осуществляется блоком ССА в процессе обработки ПС-текста и построения семантической структуры [5].

7 Особенности параметрической настройки

Опираясь на опыт построения ЛП для различных предметных областей (п. 2), чтобы постоянно учитывать все новые особенности текстовой типологии, в блок ЛМА были введены средства управления лексико-морфологическим анализом, названные средствами **параметрической настройки**. Эти средства относятся к ЛЗ и размещаются в отдельном файле. Они имеют вид списков, оформленных в виде фрагментов РСС со своими именами. Имена играют роль операторов и определяют вид анализа. Рассмотрим некоторые из них, разделив операторы на смысловые группы.

7.1. Средства идентификации начала и конца предложения.

Если слово, указанное во фрагменте NEW_SENT, записано в тексте с прописной буквы и находится в начале строки, то оно рассматривается как начало нового предложения.

Если в тексте встречается одно из слов (символов, знаков), указанных во фрагменте END_SENT, то оно считается концом предложения.

Фрагмент ABBR задает список сокращений с точками на конце, которые считаются цельными словами, и точки не рассматриваются как конец предложения.

Фрагмент SEPARATOR задает символы, которые всегда являются разделителями слов.

7.2. Средства для замены или удаления некорректных символов или слов.

Фрагменты LETTER_CH и WORD_BAD задают замены (или удаление) нежелательных слов или знаков в тексте.

Фрагменты BEG_SYMB задают набор удаляемых знаков в начале слова, а END_SYMB – в конце.

7.3. Средства унификации и синонимичных замен.

Фрагмент SYNON задает список синонимичных слов, которые заменяются на слово из первой позиции.

Фрагмент TERMIN_ заменяет слова, записанные на второй и последующих позициях, на слово в первой позиции.

Фрагмент SIGN_MANY задает повторяющиеся символы, следующие один за другим (например, набор черточек) на один символ (черточку).

7.4. Средства настройки морфологического анализатора.

Фрагмент MORF определяет генерацию морфологических признаков слова в виде фрагментов ПС-текста.

Фрагмент NOMO задает список слов, для которых устанавливается запрет на нормализацию и морфологический анализ.

Это необходимый набор операторов, без которых (как оказалось) трудно обеспечить качественный лексико-морфологический анализ многих текстов ЕЯ, и следовательно, качественную работу всего объектно-ориентированного ЛП.

Заключение

В данной статье рассмотрены направления развития блока лексико-морфологического анализа, используемого в объектно-ориентированных лингвистических процессорах (ЛП) при формализации текстов ЕЯ, т.е. для извлечения из них информационных объектов, признаков и связей. В блок введены дополнительные средства, с помощью которых обеспечивается устойчивая и качественная работа ЛП при обработке массивов документов на ЕЯ в различных предметных областях: «Криминалистика», «Резюме», «Терроризм», «Памятники культуры» и др.

Литература

1. Кузнецов И.П. Семантические представления / Кузнецов И.П. – М. : Наука, 1986. – 290 с.
2. Кузнецов И.П. Продукционный язык программирования ДЕKL / И.П. Кузнецов, М.М. Шарнин // Система обработки декларативных структур знаний Деклар-2. – М. : ИПИ РАН, 1988.
3. Кузнецов И.П. Семантико-ориентированные системы на основе баз знаний : [монография] / И.П. Кузнецов, А.Г. Мацкевич. – М. : МТУСИ, 2007. – 173 с.
4. Сомин Н.В. Система морфологического анализа: опыт эксплуатации и модификации / Н.В. Сомин, Н.С. Соловьева, М.М. Шарнин // Системы и средства информатики. – 2005. – Вып. 15. – С. 20-30.

5. Кузнецов И.П. Средства настройки семантико-ориентированного лингвистического процессора на выделение и поиск объектов / И.П. Кузнецов, Н.В. Сомин // Сб. ИПИ РАН. – 2008. – Вып. 18. – С. 119-143.
6. Кузнецов И.П. Принципы организации объектно-ориентированных систем обработки неформализованной информации / И.П. Кузнецов, Е.Б. Козеренко, А.Г. Мацкевич // Искусственный интеллект. – 2010. – № 3. – С. 227-237.
7. Kuznetsov Igor. The system for extracting semantic information from natural language texts / Igor Kuznetsov, Elena Kozerenko // Proceeding of International Conference on Machine Learning. MLMTA-03, Las Vegas US, 23 – 26 June 2003 г. – P. 75-80.
8. Коваленко А. Вероятностный морфологический анализатор русского и украинского языков [Электронный ресурс] / А. Коваленко. – Режим доступа : <http://www.keva.ru/stemka/stemka.html>.
9. Сегалович И. Русский морфологический анализ и синтез с генерацией моделей словоизменения для не описанных в словаре слов [Электронный ресурс] / И. Сегалович, М. Маслов // Диалог'98. – Казань : ООО «Хэтер», 1998. – Режим доступа : <http://download.yandex.ru/company/DLG98-MM2.pdf>.

Literatura

1. Kuznetsov I.P. Semanticheskie predstavlenija. M.: Nauka. 1986. 290 s.
2. Kuznetsov I.P. Produkcionnyj jazyk programirovanija DEKL. Sb. Sistema obrabotki deklarativnyh struktur znanij Deklar-2. IPI RAN. 1988.
3. Somin N.V. Sistemy i sredstva informatiki. 2005. Vyp. 15. S. 20-30.
4. Kuznetsov I.P. Sb. IPI RAN. 2008. Vyp. 18. S. 119-143.
5. Kuznetsov I.P. Iskustvennyj intellect. 2010. № 3. S. 227-237.
6. Kuznetsov I.P. Semantiko-orientirovannye sistemy na osnove baz znanij : [monografija]. M.: MTUSI. 2007. 173 s.
7. Kuznetsov I.P. Proceeding of International Conference on Machine Learning. MLMTA-03. Las Vegas US. 23 – 26 June 2003. P. 75-80.
8. Kovalenko A. Veroyatnostnyj morfologicheskij analizator russkogo i ukrainskogo jazykov. <http://www.keva.ru/stemka/stemka.html>.
9. Segalovich I. Russkij morfologicheskij analiz i sintez s generaciej modelej slovoizmenenija dlja ne opisannyh v slovare slov. Dialog'98. Kazan': ООО "Hjeter".1998. <http://download.yandex.ru/company/DLG98-MM2.pdf>.

И.П. Кузнецов, Н.В. Сомин, Е.Б. Козеренко, А.Г. Мацкевич

Особливості лексико-морфологічного аналізу в задачах добування структур знань з текстів природної мови

Розглядається клас об'єктно-орієнтованих лінгвістичних процесорів, які виділяють структури знань з текстів природної мови (ПМ). Важливою компонентою таких систем є блок лексико-морфологічного аналізу. У процесі розробки застосувань цей блок постійно удосконалювався і набув багато нових функцій, які виходять за межі можливостей існуючих блоків подібного типу. Даний блок генерує лексичні, морфологічні, семантичні ознаки слів, визначає найпростіші форми природної мови, має спеціальні засоби настройки на предметну область і на особливості текстів ПМ. У роботі розглядаються ці функції.

I.P. Kuznetsov, N.V. Somin, E.B. Kozerenko, A.G. Matskevich

Features of Lexical-Grammatical Analysis for Knowledge Retrieval from Texts in Natural Language

The paper analyses the experience of using the linguistic processor, which retrieves knowledge (information objects or essences and their links) from natural language texts. Significant part of the processor is the procedure of lexical-grammatical analysis, which has been modified in process of tuning to various subject fields. Now the procedure has many peculiarities, which are considered in the paper. The procedure generates lexical, morphological and semantic word attributes. It analyses some forms of natural language. It has special means of tuning to subject fields and to text features. These functions play a significant role in enhancing the quality of the linguistic processor.

Статья поступила в редакцию 31.05.2011.