МЕТОДИКИ ВЫЯВЛЕНИЯ ОБЪЕКТОВ И СВЯЗЕЙ, ЗАДАННЫХ В НЕЯВНОМ ВИДЕ

И. П. Кузнецов (igor-kuz@mtu-net.ru)

Институт проблем информатики РАН, Москва, Российская Федерация.

Рассматривается семантико-ориентированный лингвистический процессор, извлекающий из текстов естественного языка структуры знаний: информационные объекты (именованные сущности), их свойства, связи и участие в действиях. Одно из направлений развития таких процессоров связано с выявлением имплицитной информации, которая рассматривается в узком плане — как выявление новых свойств объектов и связей, заданных в неявном виде. Предлагаются методики такого выявления, осуществляемого в процессе синтактико-семантического анализа

Ключевые слова: извлечение знаний из текстов, лингвистические процессоры, имплицитная информация, обработка структур знаний.

THE METHODS OF DISCOVERY OF OBJECTS AND THEIR LINKS PRESENTED IMPLICITLY IN TEXTS

I. P. Kuznetsov (igor-kuz@mtu-net.ru)

Russian Academy of Science, Institute for Informatics Problems, Moscow, Russian Federation.

The paper presents the development of the semantics-oriented linguistic processor which analyzes natural language texts and extracts knowledge structures: information objects (named entities), their properties, links and participation in the actions. The methods of extracting new objects and links presented in texts in implicit forms are proposed. The methods employ the Knowledge Base technologies and consist in the transformation of knowledge structures in the process of syntactical-semantic analysis.

Key words: Knowledge extraction, linguistic processor, implicit information, knowledge structure processing.

Введение

В настоящее время проблема извлечения знаний приобретает все большую актуальность [1]. Одно из направлений связано с извлечением из текстов естественного языка (ЕЯ), так называемых, информационных объектов (лиц, организаций, адресов, дат и др.) и связей между ними (другое название объектов — «именованные сущности»). На этой основе разработаны системы «Криминал», «Аналитик» (ИПИ РАН), «Аналитический курьер» (Ай-Теко), «Semantix» (Синергетические системы), «PullEnti» и др. [1,6]. Успешность систем зависит от извлекаемой информации (количества и типов извлекаемых объектов и связей), а также от способа представления и средств обработки знаний, что непосредственно определяет класс и качество решаемых задач. Имеются в виду задачи идентификации объектов, выявления и анализа фактографической информации, семантического поиска, экспертных решений, ответа на запросы, выраженные на ЕЯ, и др. [2–4].

При извлечении знаний следует учитывать, что в текстах ЕЯ много полезной информации дается в скрытом или неявном виде. Такая информация называется имплицитной [5]. Ее тоже нужно извлекать и использовать. Данная статья посвящена этой проблеме, которая рассматривается применительно к выделению объектов и связей, и является продолжением исследований по тематике Лингво-ИИ [2].

Для извлечения знаний требуется разработка соответствующих лингвистических процессоров, отображающих тексты ЕЯ на структуры знаний. При этом формализмы представления знаний должны учитывать высокую степень разнообразия объектов и их связей. Например, для лиц должны быть представлены не только родственные связи и их анкетные данные, но и действия или события, в которых эти лица участвуют. Собственно, они и составляют факты. Такие действия привязаны к времени, месту. Более того, одни события могут быть составной частью других. Они могут быть связаны причинно-следственными и временными отношениями. Для ряда задач подобные связи играют важную роль. Их тоже нужно выявлять и использовать. Поэтому следует считать, что действия и соответствующие факты — это тоже информационные объекты, связанные между собой и с другими информационными объектами. Возникают сложные структуры знаний.

Для представления структур знаний в рамках проектов ИПИ РАН разработан язык расширенных семантических сетей (РСС), а для обработки — продукционный язык ДЕКЛ [4,6]. Они образуют образует законченный технологический комплекс, ориентированный на сложные задачи, связанные с логическим выводом, преобразованием представлений, экспертными решениями.

На той основе разработан и постоянно совершенствуется семантико-ориентированный лингвистический процессор (ЛП), анализирующий тексты ЕЯ и извлекающий из них структуры знаний — так называемые содержательные портреты документов (СП-документов) [3,4]. Они представляются в виде РСС и образуют Базу знаний (БЗ), в рамках которой обеспечивается анализ высокой степени глубины и сложности.

Отметим, что первые такие процессоры были разработаны для системы «Криминал», предназначенной для решения логико-аналитических задач ГУВД г. Москвы. Система проводит глубинный анализ документов, циркулирующих в ГУВД. выделяет до 40 типов объектов, их свойств, отношений и их участие в действиях. Система «Криминал» отлаживалась на 500 тыс. происшествий из сводок ГУВД г. Москвы. По основным объектам удалось добиться хороших результатов: коэффициент шумов в компонентах (лишних слов в объектах) — не более 1–2 % и потерь (отсутствие нужных слов) — не более 1 % [3, 4].

В данной статье рассматривается развитие таких процессоров (ЛП), связанное с извлечением из текстов ЕЯ имплицитной информации. Рассматриваются методики извлечения объектов («сущностей») и связей, заданных в неявном виде.

1. Средства представления и обработки знаний

С помощью семантико-ориентированного ЛП из текстов ЕЯ извлекаются информационные объекты и связи, а также конструкции ЕЯ, представляющие связи, действия (факты). Они преобразуются в однотипные фрагменты на РСС, имеющие вид:

<тип объекта>(<арг.1>,<арг.2>,…/<код фрагмента>),
<вид связи>(<арг.1>,<арг.2>,…/<код фрагмента>),
<имя действия>(<арг.1>,<арг.2>,…/<код фрагмента>).

Код фрагмента — это константа, которая соответствует объекту или действию, представленному с помощью всего фрагмента. Аргументами (арг. N) могут быть слова в нормальной форме (необходимо для идентификации и поиска), или коды других фрагментов. В результате обеспечивается представление случаев, когда одни объекты включают в себя другие, или когда комплексные действия включают в себя объекты и другие действия. Такие случаи недопустимы в логике предикатов, но являются типичными для текстов ЕЯ, что легко представляется в виде РСС, и соответственно, в БЗ..

Отметим, что вся информация представляется в БЗ на однородной основе, что очень важно для обработки, осуществляемой продукциями языка ДЕКЛ. Левая и правая части таких продукции (правила ЕСЛИ …ТО) состоят из аналогичных фрагментов, содержащих переменные. Последние означиваются в процессе применения продукций — сопоставления ее левой части со структурами в БЗ и выполнения действий, указанных в правой части. Как показывает опыт, РСС и ДЕКЛ составляю универсальную инструментальную среду, ориентированную на представление и обработку семантической информации, извлекаемой из текстов ЕЯ.

Процессор ЛП реализован средствами языка ДЕКЛ и управляется лингвистическими знаниями (ЛЗ) в виде предметных словарей, средств параметрической настройки, а также правил выделения объектов и связей [3,4].

С помощью ЛЗ осуществляется настройка ЛП на соответствующие категории пользователей и корпуса текстов. В результате возникает конкретная реализация. Таким образом, речь идет о средствах построения класса процессоров с широкими возможностями их настройки и совершенствования.

2. Принципы выявления новых объектов и связей

Для выявления многих объектов используются характеристические слова, по которым определяется наличие объекта. Например, слова «дом» (за которым стоит число) или «улица» (за которым стоит слово с большой буквы) определяют наличие объекта типа «адрес». Аналогично, слова «фирма», ООО, «банк» и др. (за которыми стоит слово с большой буквы или слова в кавычках) определяют наличие объекта типа «адрес». Это характеристические слова, с которых начинается выделение объекта, включающего эти слова.

При отсутствии характеристических слов используется принцип ожидания — после одних слов или объектов ожидается наличие других. Например, если после слова «инженер» стоит слово с большой буквы, то скорее всего, оно относится к ФИО. Вместо слова «инженер» может быть любое другое слово, выражающее профессию. При этом нужно учитывать наличие между этим словом и ФИО факультативных элементов, например, названия организации. Таким образом начинается выделение подразумеваемых объектов, т. е. у которых нет характеристических слов, определяющих их наличие. Например, не распознаны компоненты ФИО.

В текстах ЕЯ многие связи подразумеваются и привязаны к типу выявленных объектов. Например, если выявлен адрес, то скорее всего, он относится к какому-либо определенному лицу (или организации), которое нужно искать. При результативном поиске формируется новая связь. На этом основана методика формирования новых связей. Она заключается в следующем. В процессе анализа текста строятся «временные» фрагменты, представляющие связи выявленных объектов с пока что неизвестными объектами, которые специальным образом отмечаются. В дальнейшем осуществляется их поиск. Если соответствующий объект не найден, то «временный» фрагмент удаляется из СП-документа. Если найден, то фрагмент остается и вводится в структуру СП-документа.

Аналогичная методика используется при формировании новых признаков. Формируется признак с пока что неизвестным объектом, который в дальнейшем уточняется.

При формировании объектов некоторые компоненты могут быть сразу не найдены, например, год рождения, который в СП-документа представляется как компонента ФИО. Тогда в соответствующих фрагментах специальными константами отмечаются незаполненные аргументные места, которые в дальнейшем уточняются. Для более детального описания методик и средств их реализации рассмотрим правила и этапы построения СП-документов в процессе синтактико-семантического анализа.

3. Правила синтактико-семантического анализа

Синтактико-семантический анализ необходим для выделения связанных групп слов, а также информационных объектов: адресов, номеров машин, организаций и др. Последние, как правило, это наборы слов, которые могут быть грамматически никак не согласованы. Их выделение осуществляется по чисто формальным принципам на основе правил, составляющих ЛЗ. Например, адрес может рассматриваться как набор буквосочетаний «г.», «ул.», «д.»,.., слов с большой буквы и чисел. Каждый такой набор может иметь свои границы и недопустимые компоненты. Например, в адресах не может быть местоимений, глаголов и т. д. Выделение таких наборов слов, составляющих описания объектов, основано на использовании правил синтактико-семантического анализа (в дальнейшем просто — правил) следующего вида:

<ПравилоN>:CONTEXT(<слово1>,<слово2>, ...) --> <результ. фрагмент>

где <ПравилоN> — имя правила, необходимое для его вызова, а <слово1>,<слово2>,

... — это может быть отдельное слово, признак, а также И-ИЛИ граф, составленный из слов и признаков. Для этих правил указывается, с какой позиции начинать применение, а также допустимый или недопустимый контекст. Обычно применение начинается с позиции, на которой находятся характеристические слова. Например, выделение лиц начинается с поиска распознанных компонент ФИО. Выделение адресов с поиска слов ул., дом, кв. и т. д.

Правила выделяют из текста группы слов (по их признакам), описывающих какой-либо объект, и заменяют их на одно (абстрактное) слово, с которым связывается соответствующий фрагмент семантической сети и которому присваиваются определенные признаки (см. ниже), в том числе признак, указывающий на тип объекта.

Синтактико-семантический анализ предложений с выделением словосочетаний и анализом форм осуществляется на основе правил, которые применяются в определенной последовательности. Вначале выделяются простейшие объекты, затем -согласованные группы слов, затем — более сложные объекты и их признаки, и наконец, глагольные формы, см. п. 4. По мере применения таких правил строится семантическая сеть — содержательный портрет документа. Например, рассмотрим правило с именем GG~1:

MUSTBE(GG~1,1) STR_OR(ADJ,PRON/2+) CONTEXT(2-,NOUN/GG~1)
P_P(GG~1,3+) WORD_C(1,2/3-) NOTBE(GG~1,2,LETT)
Правило GG~1 осуществляет преобразования:
GG~1:ПРИЛАГАТЕЛЬНОЕ + СУЩЕСТВИТЕЛЬНОЕ --> <комбинация слов>
МЕСТОИМЕНИЕ + СУЩЕСТВИТЕЛЬНОЕ --> <комбинация слов>.

Фрагмент MUSTBE указывает, что применять правило GG~1 нужно с 1-ой позиции, т. е. искать слова с признаками ПРИЛАГАТЕЛЬНОЕ (ADJ)

и МЕСТОИМЕНИЕ (PRON), так как их меньше, чем СУЩЕСТВИТЕЛЬНЫХ (NOUN). Символ 2+ это код фрагмента типа «ИЛИ» (STR_OR), а фрагмент CONTEXT(2-,NOUN/GG~1) задает позиции правила GG~1, где на первой позиции стоит указанный код (его повторное применение обозначается 2-), а на второй — признак NOUN. Аналогичным образом используются символы 3+ и 3-.

Фрагмент P_P отделяет левую часть от правой (- ->), а WORD_C — указывает, что слова на 1-й и 2-ой позициях должны быть склеены в комбинацию слов, которое в дальнейшем будет рассматриваться как одно слово с морфологическими признаками 2-го слова. Фрагмент NOTBE указывает, что на 2-ой позиции не могут быть отдельные буквы (признак LETT). К данному правилу добавляется фрагмент, требующий согласованности слов (по падежам, числам), а также фрагменты, задающие с признаков и контекстные ограничения.

Это пример наиболее простого правила. Помимо правил выделения словосочетаний и объектов, в ЛЗ имеются специальные правила, которые осуществляют идентификацию объектов, например, с местоимениями или краткими описаниями (по имени восстанавливается фамилия, если они где-нибудь упоминались вместе). И многое другое, что необходимо для работы с текстами ЕЯ.

Отметим, что каждое правило (как и все лингвистические знания) записывается на языке РСС и является частью ЛЗ. Над правилами работают продукции языка ДЕКЛ (программа), которые применяют эти правила и играют роль пустой лингвистической оболочки, поддерживающей язык записи лингвистических знаний — РСС. Как показывает опыт, такую оболочку можно настраивать на различные языки, т. е. строить различные лингвистические процессоры, в том числе, англоязычные [6].

4. Порядок применения правил

Правила синтактико-семантического анализа применяются в строго определенной последовательности — каждое на своем уровне. Например, при обработке сводок происшествий вначале выделяются информационные объекты — отделения милиции (ОВД_), сотрудники милиции (МИЛ_) и др. Они могут содержать фамилии, имена, которые следует отличать от ФИО лиц — фигурантов (последние представляются фрагментами FIO). Далее выделяются статьи УК и т. д. Это необходимо, чтобы облегчить последующий анализ. Иначе слова, составляющие эти объекты, могут захватываться другими правилами и создавать шумы.

Далее начинается выделение лиц — фигурантов. Для этого вводится множество правил. Одни правила начинают свое применение с поиска распознанных имен или фамилий (MUSTBE), другие — с поиска года рождения, третьи — с инициалов. В результате минимизируются потери в случаях, когда блок морфологического анализа не дает необходимых признаков для каких-либо слов (что это имена или фамилии и т. д.).

Затем анализируются словосочетания, выделяются объекты, и наконец, анализируются глагольные формы. По мере применения таких правил

строится СП-документа. Последовательность правил задается с помощью специальных фрагментов. Ниже приведен пример представления уровней, определяющих порядок применения правил.

```
{== Уровни ==}
LEVEL(LEVEL1,LEVEL2,LEVEL3,LEVEL4,...)
LEVEL1(CATALOG)          {= Объединение словосочетаний из каталогов =}
. . .
LEVEL2(MIL~1,ST~1)                    {= Выявление отд.милиции, ст. УК =}
LEVEL3(DD~1,DD~2,…)  {= Выявление времени, дат, в том числе, г.рожд. =}
LEVEL4(FF~1,FF~2, …)          {= Выявление лиц с распознанными ФИО =}
LEVEL4(FA~1)                        {= Выявление нераспознанных лиц =}
LEVEL4(ID_4)                {= Поиск года рождения для выявленных лиц =}
LEVEL4(PROP~1,PROP~2,ID_33        {= Выявление свойств и поиск лиц =}
. . .
LEVEL5(AA~1,AA~2)                     {= Выявление однородных членов =}
LEVEL6(GG~1,GG~2,...)                    {= Выявление словосочетаний =}
. . .
LEVEL10(ID_1)                 {= идентификация связок «тот, который» =}
LEVEL11(ID_2A,ID_2,ID_21)              {= идентификация местоимений =}
. . .
```

В фигурных скобках даны комментарии. Первый фрагмент LEVEL(…) задает уровни, а последующие — правила каждого уровня.

Правила начинают применяться к семантической сети, которая имеет вид линейной структуры и которая представляет последовательность слов в нормальной форме. Такая сеть формируется блоком лексико-морфологического анализа [8]. При этом последовательность слов задается с помощью фрагментов LR, с которыми связываются распознанные признаки слов: лексические, морфологические, семантические. Предложения разделяются фрагментами SENT. Все это представляется на РСС.

Правила анализируют линейную структуру, находят соответствующие группы слов, из которых формируются объекты. При этом объекты как бы замещают эти слова. Линейная структура сохраняется, но видоизменяется. В конце остается линейная структура (на РСС), компонентами которой являются объекты и слова, не вошедшие в объекты (напомним, что события и действия — это тоже объекты). На этой основе формируется СП-документа [5,6] .

В ЛП имеются правила, которые обеспечивают полный разбор предложений. При этом параллельно обеспечивается выделение значимых (информационных) объектов, в том числе таких, в которых слова никак не согласованы между собой, например, адресов, машин с указанием их номеров и т. д. [3,4].

5. Принцип «ожидания» при выявлении объектов

При наличии в тексте объектов без характеристических слов возникают трудности их выделения. Например, если в тексте встречаются лица с иностранными ФИО. У английских фамилий («Арафат», «Райс», «Браун», …) нет характерных суффиксов, как в русском языке. Более того, в качестве фамилий может быть любое слово, называющее или определяющее какой-либо предмет внешнего мира. При анализе англоязычных текстов такие фамилии вносят элементы неопределенности — омонимии. В азиатских языках компоненты ФИО — это просто слова с большой буквы («Ден Сяо Пин», «Лю Шао Ци», …). Задать перечислением имена или фамилии (в предметном словаре) не представляется возможным. В таких ФИО отсутствуют характеристические слова. Требуются другие методики выделения. Аналогично, адреса могут иметь вид — «Никольская 12–55». Сказанное относится и к другим объектам.

Для выделения, как уже говорилось, используется принцип «ожидания» — после одних объектов (или понятий) ожидается наличие других. Реализация соответствующей методики осуществляется с помощью операторов вида:

GO_(<Правило1>,<Правило2>,N),

где Правило1 — правило, которое было вызвано. И если оно применилось, то оно вызывает Правило2, применение которого начинается с позиции N.

Рассмотрим пример использования данного оператора при выявлении ФИО. Это осуществляется с помощью двух правил — FA~1 и FF~1:

```
MUSTBE(FA~1,1) STR_OR(WORK_K,NAT_K/2+) CONTEXT(2-/FA~1)
P_P(FA~1," ") GO_(FA~1,FF~1,1)

MUSTBE(FF~1,1) STR_OR(NAME0/3+) CONTEXT(3-,3-,3-/FF~1)
P_P(FF~1,4+) FIO(1,2,3," "/4-) MAYBE(FF~1,3)
STR_OR(VERB,ENG/5+) NOTBE(FF~1,ALL,5-)
```

Правило FA~1 находит в тексте слова с признаками WORK_K (профессии) и NAT_K (национальность). Такие признаки присваиваются словам блоком морфологического анализа на основе предметных словарей, где даны списки профессий, национальностей и др. [9,10]. И если слово с таким признаком найдено, то вызывается правило FF~1, которое проверяет, чтоб за найденным словом стояли 3 слова с большой буквы (с признаком NAME0). При этом такие слова не могут быть (NOTBE) глаголами (которые имеют признак VERB) или англоязычными (их признак — ENG), что задается с помощью двух последних фрагментов. Фрагмент MAYBE(FF~1,3) указывает, что третья позиция является факультативной, т. е. третьего слова с большой буквы (ББ) может не быть. И все одно правило будет применимым. В случае применимости формируется фрагмент FIO(…). У него в качестве первых трех аргументов будут первые три слова, которые удовлетворяют условиям, заданным в фрагменте CONTEXT. Эти три

слова заменяются на одно, с которым связывается сформированный фрагмент и к которому добавляется признак FIO.

Эти два правила осуществляют преобразования:

ПРОФЕССИЯ + 2 или 3 СЛОВА С ББ --> <выделенное лицо>
НАЦИОНАЛЬНОСТЬ + 2 или 3 СЛОВА С ББ --> <выделенное лицо>

Например, словосочетание «председатель Ху Цзинь Тао» будет преобразовано в фрагмент FIO(ХУ,ЦЗИНЬ,ТАО,” ”). При этом слово «председатель» останется и будет использовано при последующем анализе. Словосочетание «премьер Хапер Стивен» будет преобразовано в фрагмент FIO(ХАПЕР,СТИВЕН,” ”,” ”). Для выделения ФИО из словосочетаний типа «премьер Канады Хапер Стивен» в фрагмент CONTEXT первого правила необходимо вставить факультативную позицию для слов с признаком «государство». Путем модификации правил можно охватить множество случаев не увеличивая количество правил.

Другой способ выделения ФИО — через глаголы, субъектами которых могут быть только лица. Например, «… Хапер Стивен подписал …», где глагол «подписать» помогает выделению лица. Такие глаголы даются перечнем («предложить», «подписать», «согласиться», …), а выделение лиц реализуется с помощью того же оператора GO_.

6. Выявление признаков и связей

В данном разделе рассматривается методика выявления связей, заданных в неявном виде. Для этого в правые части синтактико-семантических правил, выявляющих определенного типа объекты, вводятся «временные» фрагменты, представляющие связь этих объектов с пока что неизвестными объектами, которые в дальнейшем ищутся и уточняются с помощью специальных процедур идентификации. Если неизвестный объект найден, то «временный» фрагмент становится постоянным и вводится в структуру СП-документа. Например, для адреса строится фрагмент ИМЕТЬ(??_1,<адрес>) и в дальнейшем с помощью процедур идентификации осуществляется поиск аргумента ??_1, соответствующего лицу или организации. Найденный объект замещает этот аргумент.

Другой вариант имеет место, когда предполагается, что у лица, встретившегося в тексте, должен быть задан адрес. Тогда в правую часть правила, выявляющего лица, вставляется другой фрагмент ИМЕТЬ(<лицо>, ??_2), где аргумент ??_2 соответствует адресу. В дальнейшем осуществляется его поиск.

Выбор варианта зависит от вероятности наличия связи, что определяется особенностью корпусов анализируемых текстов. Например, в сводках происшествий не для каждого человека может быть задан адрес. И в тоже время, если встретился адрес, то он, как правило, относится к какому-либо лицу. Реже — к организации. И очень редко — к действию или событию. В корпусах текстов области «Резюме», где описываются данные людей для приема на работу, наоборот. Человек, который пишет резюме, должен указать свой адрес, телефон

и т. д. Поэтому и правила, составляющие ЛЗ для каждой области будут иметь свои особенности.

Рассмотрим одно из таких правил, соотносящих клички к лицам — фигурантам.

```
MUSTBE(FFA~1,2)
STR_OR(КЛИЧКА,ПСЕВДОНИМ/1+)
STR_OR(NAME0,КВЧ/2+)
CONTEXT(1-,2-/FFA~1) КЛИЧКА(??_1,2/3+) P_P(FFA~1,3-)
GO_(FFA~1,ID_33)
```

Данное правило FFA~1 ищет словосочетания следующего вида:

кличка или псевдоним + <слово с большой буквы (NAME0) или слово в кавычках (КВЧ)>.

И если такое словосочетание найдено, то формируется фрагмент КЛИЧКА(??_1,<2-е слово>), где аргумент ??_1 соответствует неизвестному лицу. После этого с помощью оператора GO_(FFA~1,ID_33) вызывается процедура идентификации ID_33, которая осуществляет поиск лица. В результате формируется законченный фрагмент.

Например, если анализируется текст «… Агджа Мехмет Али 1945 г. р. …, скрывался под псевдонимом Хаджи …» , то правило FFA~1 делается применимым к последнему словосочетанию. В результате формируется фрагмент КЛИЧКА(??_1,ХАДЖИ).

После этого с помощью оператора GO_ вызывается процедура ID_33, которая осуществляет поиск лица. Код фрагмента, соответствующего найденному лицу, подставляется на место аргумента ??_1. В результате в СП-документа формируются связанные фрагменты:

```
FIO(АГДЖА,МЕХМЕТ,АЛИ.1945/3+) КЛИЧКА(3-,ХАДЖИ).
```

Отметим, в другом варианте в правила, осуществляющие поиск лиц, могут быть вставлены фрагменты, связывающие лица с пока что неизвестными кличками. Но вероятность такой связи не велика, что делает последующие поиски кличек мало результативными.

Процедура идентификации неизвестных объектов задается в ЛЗ с помощью специальных правил идентификации, каждое из которых содержит фрагмент ID_K, где указывается, какого типа объекты следует искать, в каком направлении и когда заканчивать поиск. Поиск заключается в последовательном переходе по шагам от одного компонента линейной структуры (слова, выявленного словосочетания или объекта) к другой, начиная от того места, где встретился знак неизвестного объекта — ??_N.

Отметим, что правила идентификации могут вызываться на любом уровне анализа текста, см. п. 4 (а не только с помощью операторов GO_). Важно,

чтобы при вызове правила объекты, которые оно должно искать, были бы уже выявлены.

Фрагмент ID_N имеет следующую структуру:

ID_K(??_N,A3,A1,A2,LEFT),

где ID_K — имя правила (через него осуществляется вызов);
??_N — указывает на неизвестный объект;
A1 — задает тип объекта, который нужно искать;
LEFT — указывает, что искать объект нужно слева (RIGHT — справа);
A2 — ограничивает количество шагов поиска;
A3 — задает поисковый режим: заканчивать (или нет) поиск, если встретился символ конца (или начала) предложения.

Поиск начинается от того места линейной структуры, где встретился знак неизвестного объекта — ??_N. И заканчивается, если найден нужный объект или выполнены условия окончания. Это может быть: допустимое количество шагов, наличие символа начала предложения, а также специальные условия. Для их представления к основному фрагменту ID_K добавляются фрагменты, которые задают недопустимые слова — в виде списка ли перечня:

STR_OR(<перечень недопустимых слов и признаков>/2+) NOTBE(ID_32," ",2-)

Если в процессе движения по линейной структуре встретилось слово, входящее в перечень, или компонента (слово, словосочетание, объект), имеющее признак из перечня, то движение заканчивается. Поиск считается не результативным.

Рассмотрим пример поиска неизвестных лиц, отмеченных символом ??_1.

STR_OR(LR,SENT/1+) {= Допускается переход по словам и по предложениям =}
STR_OR(FIO/2+) {= Определяет идентификацию ??_1 — с лицами =}
STR_OR(ЗАДЕРЖАТЬ,НАНЕСТИ,POINT_1/2+) {= Что не допустимо при переходах =}
ID_33(??_1,1-,2-,20,LEFT) NOTBE(ID_33," ",2-)

Это набор управляющих фрагментов, где основным является — ID_33. Этот фрагмент определяет движение влево (LEFT) по линейной структуре от того места, где находится знак ??_1, с поиском фрагмента FIO(…), представляющего лицо. Число шагов поиска — не более 20. При этом допускается переход по словам (LR), а также от одного предложения к другому (SENT), т. е. поиск не заканчивается, если встретился символ начала предложения (или конца предыдущего). Но поиск заканчивается, если встретились глаголы «задержать», «нанести» или слово с признаком POINT_1 — это числа с точкой в конце, стоящие вначале строки.

Аналогичная методика используется при формировании новых признаков объектов. Рассмотрим пример:

```
MUSTBE(PROP~2,2)
STR_OR(БЕЗРАБОТНЫЙ,ПОТЕРПЕВШИЙ,ЗАЯВИТЕЛЬ/1+)
CONTEXT(1-/PROP~2) 1(??_1/2+) P_P(PROP~2,2-)
GO_(PROP~2,ID_33) {== Уточняется ??_1 (чье свойство) ==}
```

Правило PROP~2 ищет слова «безработный», «наркоман», «преступник» (их может быть больше). И если, к примеру найдено слово «безработный», то на основе 1(??_1/2+) формируется фрагмент типа ПОТЕРПЕВШИЙ(??_1). Далее вызывается (GO_) правило идентификации ID_33, которое ищет лицо, к которому относится данное свойство. Конечно, правила выявления лиц должны быть вызваны раньше, чем PROP~2. Отметим, что вызов правила ID_33 можно осуществлять через уровни обработки (см. п.4). Однако, оператор GO_ делает такой вызов более целенаправленным.

7. Уточнение неопределенных компонент

Достаточно часто при анализе текста, выявлении объектов и формировании соответствующих фрагментов РСС некоторые компоненты могут оставаться неизвестными. Например, если они описаны где-то в другом месте. Например, в текстах резюме год рождения может находиться на значительном расстоянии от лица. В сводках происшествий имеет место тот же случай. Например, «… Сидоров Иван, ведущий инженер ООО «Вымпел», … 1966 г. р., …». Тогда формируется фрагмент с неизвестным компонентом FIO(СИДОРОВ,ИВАН,” ”,??_2), где аргумент ??_2 в дальнейшем уточняется с помощью соответствующего правила идентификации.

Рассмотрим пример.

```
MUSTBE(FF~1,2)
CONTEXT(FAM,NAME,NAME_1/FF~1)
FIO(1,2,3,??_2/3+) P_P(FF~1,3-) 3-(FIO,ADD_)
```

С помощью правила FF~1 осуществляется поиск трех слов, где первое имеет признак фамилия (FAM), второе — имя (NAME), а третье — отчество (NAME_1). Это задается фрагментом CONTEXT(…/FF~3). Фрагмент MUSTBE(FF~1,3) указывает, что применять правило нужно с 2-ой позиции, т. е. с поиска распознанных имен (типичные русские имена даются в предметном словаре). Если поиск оказался результативным, то в рамках линейной структуры формируется фрагмент, который замещает эти слова:

```
FIO(<1-е слово>,<2-е слово>,<3-е слово>,??_2).
```

Далее на уровне LEVEL4(ID_4), когда уже сработали правила выделения дат и года рождения (см. п.4), вызывается правило идентификации ID_4 следующего вида:

```
STR_OR(LR,SENT/25+)
STR_OR("год рождения"/26+) {== С чем идентифицируется ??_2 ==}
ID_4(??_2,25-,26-,8,RIGHT) {= Ищет вправо от фрагмента с ??_2 =}
```

Это управляющие фрагменты, указывающие на необходимость поиска фрагментов с аргументом ??_2 и перемещения от каждого фрагмента вправо (RIGHT) с поиском объекта (или числа) с признаком "год рождения". При этом задается ограничение — не более, чем 8 шагов. Этого достаточно, учитывая, что многие объекты уже найдены (описывающие их слова заменены на одно слово) и перемещение по каждому из них — это один шаг.

Отметим, что правила выделения объектов и правила идентификации представлены в лингвистических знаниях в виде наборов элементарных фрагментов РСС, которые легко менять, настраивая лингвистический процессор (ЛП) на ту или иную предметную область. Сама программа (на языке ДЕКЛ) остается неизменной. Этот фактор дает большие преимущества при отладке и настройке ЛП, так как учесть даже малую часть того, что может встретиться в ЕЯ, не представляется возможным.

8. Оценка предлагаемых методов

Предлагаемые методы реализованы в системах «Криминал» и «Аналитик». Анализ проводился на сводках происшествий и показал (после настройки лингвистических знаний) достаточно хорошие результаты. При выявлении признаков и связей количество потерь (когда связь не выявлена) не более 5%, а количество шумов (выявлении «лишних» связей) не более 1%. При уточнении неопределенных компонент количество шумов (когда компоненты неправильно означивались) около 2%, а потерь (компоненты никак не означивались) не более 6%. Эти показатели оказались приемлемыми с точки зрения решения логико-аналитических задач ГУВД г. Москвы, а также задач ГУСТМ МВД России.

Задача выделения свойств и отношений рассматривается во многих системах извлечения знаний. Следует отметить наиболее продвинутые системы, разработанные в Станфордском университете (Stanford NER system), Илинойском университете (Illinois NER system), а также “Lingpipe NER system” и др. Однако первая система (7 class) выделяет только 7 типов объектов — именованных сущностей (named entities — NE). Другие — и того меньше. Как правило, работа системы заканчивается разметкой текстов с выделение компонент, соответствующих объектам (NE) и указанием их типов.

Для решения задач логико-аналитической обработки этого недостаточно. В системах «Криминал» и «Аналитик» потребовалось выделение значительно

большего числа объектов (до 40 типов) и связей, а также выделение действий и фактов участия объектов в действиях. В результате формируются структуры знаний. Для этого разработан язык представления знаний (РСС). На этой основе создаются и реализуются новые методики и технологии, в том числе, описанные в данной статье.

Заключение

В данной статье рассмотрены семантические методики по извлечению некоторых видов имплицитной информации из текстов естественного языка. Предлагаемые методики реализованы в рамках единого инструментального комплекса (языка представления знаний РСС и обработки ДЕКЛ), ориентированного на организацию баз знаний (БЗ) и на их использование для решения интеллектуальных задач, в том числе, связанных с извлечением структур знаний, их анализом для дополнения и корректировки структур, логическим выводом, принятием экспертных решений. Предметные и лингвистические знания представляются на единой основе (в виде фрагментов РСС), что позволяет свести казалось бы разнородные задачи к преобразованию структур знаний. Это дает определенные преимущества: упрощает создание соответствующих программ (на языке ДЕКЛ), обеспечивающих анализ высокой степени глубины и сложности.

Литература

1. *Banko M., M. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni.* Open Information Extraction from the Web. Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI-07), 2007. P. 2670–2676.
2. *Kuznetsov I. P.* Identifying role functions of persons on the basis of knowledge structures. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2011" [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2011"]. Bekasovo, 2011, p. 391–402.
3. *Kozerenko E. B., Kuznetsov I. P.* The system for extracting semantic information from natural language texts. Proceeding of International Conference on Machine Learning. MLMTA-03, Las Vegas US. 2003, p. 75–80.
4. *Kuznetsov I. P., Matskevich A. G.* Semantics-oriented systems on the base of Knowledge Base (book) [Semantiko-orientirovannye sistemy na osnove Baz Znanii] Sviaz'izdat MTUSI Moscow, 2007 ,173p.
5. *Pirogova I'ui. K.* Implicit information as means of communicative influence and manipulation [Implitsitnaya infomatsiya kak sredstvo kommunikativnogo vozdeystviya I manipulirovaniya] Problemi prikladnoy lingvistiki [Problems of applied linguistics] . Moscow 2001. p. 209–227.
6. *Site* of lab. 14 IPI RAN — http://IpiranLogos.com

7. *Kuznetsov I. P., Kozerenko E. B.* Linguistic Processor “Semantix” for Knowledge extraction from natural texts in Russia and English. Proceeding of International Conference on Machine Learning, ISAT-2008. Las Vegas, USA CSREA Press, 2008, p.835–841.
8. *Somin N. V., Kuznetsov I. P.* Peculiarity of lexical-grammatical analysis for object extraction from natural language texts [Osobennosti lekciko-morfologicheskogo analiza pri izvlechenii informatsionnyx objectov i sviazei iz tekstov estestvennogo iazika]. Komp’iuternaia Lingvistika i Intellektual’nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii “Dialog 2010” [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog 20010”]. Bekasovo, 2010 p. 254–264.