СЕМАНТИКО-ОРИЕНТИРОВАННЫЙ ЛИНГВИСТИЧЕСКИЙ ПРОЦЕССОР ДЛЯ АВТОМАТИЧЕСКОЙ ФОРМАЛИЗАЦИИ АВТОБИОГРАФИЧЕСКИХ ДАННЫХ.

И.П. Кузнецов *(igor-kuz@mtu-net.ru)*

А.Г. Мацкевич

ИПИ РАН, г. Москва

Рассматриваются прямой и обратный лингвистические процессоры для обработки автобиографических данных, заявок на работу (резюме), представляющих собой тексты естественного языка. Человек в свободной форме дает сведения о себе: ФИО, год рождения, адрес, время и место учебы с указанием наименования учебного заведения, факультета, специальности, места работы (организации) с указанием периода работы, должности, основных обязанностей и др. Эти данные могут быть выражены различными способами и произвольным образом разбросаны по тексту. Задача прямого лингвистического процессора - выделение этих данных, приведение их к стандартному виду и установлению связей между ними: соотнесение организаций с датами, должностями и др. На этой основе строятся структуры знаний. Задача обратного лингвистического процессора - представление этих структур в виде компонент естественного языка (словосочетаний, предложений) и их отображение на поля формализованной анкеты или структурированного сайта.

Введение

Трудности автоматической формализации анкетных и автобиографических данных определяются, во-первых, разнообразием форм ЕЯ, с помощью которых могут выражаться одни и те же сведения. Например, даты могут быть записаны в сокращенной форме (авг.05), в виде дробных чисел (09.99 г.), разного рода специальных знаков или кавычек (09/99 или 09'1999) и т.д. Тоже самое относится к ФИО, адресам и др. Их нужно приводить к стандартному виду. Во-вторых, организации (где человек работал), должности, периоды работы и основные обязанности могут быть записаны в произвольной последовательности, что приводит к трудностям их выделения и компоновки. Если период работы в какой-либо организации записан в конце и далее идет другая организация, то нужно уметь определять, куда отнести этот период. Тем более, что периоды или даты могут находиться в разных местах, в том числе, внутри текста описания работы. Человеку по смыслу проще понять, что к чему относится. Значительно труднее выработать формальные критерии разделения и соотнесения, которые бы давали допустимое количество шумов и потерь. По указанным причинам не всегда дают эффект алгоритмы, которые делят текст на достаточно самостоятельные части и проводят анализ по частям. В-третьих, определенные трудности вызывает большое количество сокращений, ошибок, отсутствие знаков препинания (точек), наличие спецзнаков, остающихся после перекодировки текстов (для работы блока морфологического анализа), см. приложение 1.

Для формализации анкетных и автобиографических данных предлагается система LINGVO-MASTER. В ней использованы методики, разработанные в ИПИ РАН [1] и основанные на технологии баз знаний (БЗ). Особенность методик - в переносе сложных этапов лингвистического анализа на уровень обработки структур знаний, где за счет использования инструментальных средств DECL реализуются сложные виды синтактико-семантического анализа и идентификации [3]. Система LINGVO-MASTER содержит прямой и обратный лингвистические процессоры (ЛП), которые управляются с помощью лингвистических знаний (ЛЗ). ЛЗ - это также структуры знаний, хранящиеся в БЗ. ЛЗ прямого ЛП представляют собой контекстные правила специального вида, см. п.3. Прямой ЛП преобразует тексты в структуры знаний, называемые содержательными портретами документа (анкеты), где с помощью РСС представлены информационные объекты и их связи. Далее идет обработка этих знаний - для уменьшения шумов и потерь. Обратный ЛП преобразует структуры знаний в компоненты ЕЯ и отображает их на поля анкеты. ЛЗ обратного ЛП отражают вид формируемой анкеты и связь ее полей с структурами БЗ.

1. Представление знаний

Знания (предметные и лингвистические) в системе LINGVO-MASTER представляются в виде структур, которые записываются в нотации семантических сетей, дополненных средствами представления событийных компонент и комплексных связей. В результате образуются расширенные семантические сети (РСС). РСС состоит из элементарных фрагментов, имеющих произвольное количество аргументных мест (но не более 200) и представляющих свойства, отношения, события, действия. Множество фрагментов - это РСС [2].

В простейшем случае фрагмент имеет вид N-местного предиката. Например, ОРГ_(ЗАВОД,КРИСТАЛЛ) - это фрагмент, представляющий организацию. В тоже время фрагмент - это более сложная конструкция, которая далеко выходит за рамки типовых предикатов логики 1-го и 2-го порядков. Во-первых, во фрагментах широко используются внутрисистемные коды - это числа, к которым добавляется знак плюс (+), когда вводится новый код, или знак минус (-), когда используется уже введенный код. Например, "1+" и "1-" - есть обозначение одного и того же объекта (или отношения), а "2+" и "2-" - уже другого, и т.д. Такие числа служат для обозначения неименованных объектов, например, порождаемых самой системой. Во-вторых, вводится специальный код фрагмента, соответствующий всей представленной в фрагменте информации. Например, в фрагменте АДР_(УЛ.,ГЛАГОЛЕВА,25,1,273/6+) код 6+ представляет весь адрес. Эти коды могут стоять на аргументных местах других фрагментов. Например, фрагменты

FIO(ФИРСОВ,ВЛАДИМИР,НИКОЛАЕВИЧ,1953/5+)
АДР_(УЛ.,ГЛАГОЛЕВА,25,1,273/6+) ПРОЖ.(5-,6-)

представляют, что фигурант Фирсов Владимир Николаевич (ему сопоставлен код 5+, 5-) проживает (ПРОЖ.) по указанному адресу, которому сопоставлен код 6+, 6-. Коды фрагментов необходимы для представления комплексной информации и различных видов связей. РСС нашли широкое применение для представления семантической информации, содержащихся в текстах на ЕЯ (системы ДИЕС, ИКС, АНАЛИТИК). Одно и тоже понятие может называть различные объекты одного типа, которые нужно различать. Отсюда необходимость в внутрисистемных кодах. РСС ориентированы на отображение возможности интеграции множества связанных объектов в один объект, что выражается на ЕЯ в виде форм с отглагольными существительными. Понятие связи рассматривается в широком смысле. Это могут быть не только отношения, но и зависимости. Связанными считаются также объекты, участвующие в одном действии. Группа связанных объектов может быть связана с другой группой, что на ЕЯ выражается в виде глагольных форм с отглагольными существительными.

2. Содержательные портреты документов

Сеть (РСС), представляющая объекты и связи какого-либо документа (анкеты), образует так называемый содержательный портрет этого документа. Такие портреты необходимы для обеспечения быстрого и качественного поиска информации по значимым компонентам и связям. Приведем пример. Типовое неформализованное резюме рекрутингового агентства:

РЕЗЮМЕ
Ф.И.О. Евгения Александровна Иванова.
Дата рождения: 20 февраля 1977 года
Образование высшее: Ташкентский Финансовый институт, Финансово-кредитный
факультет, специальность - <Финансы предприятий различных форм собственности>.
Время обучения: 1995-2000г.г.
Трудовая деятельность:
С 1994г- начала работать бухгалтером- кассиром, материальным бухгалтером на
машиностроительном предприятии по изготовлению оборудования
"Техинпром" - Республика Узбекистан....
(полный текст резюме, см. Приложение 1)

Его содержательный портрет:

ДОК_(0,TEXT.VOP," "/0+) 0-(RUS)
FIO(ИВАНОВА,ЕВГЕНИЯ,АЛЕКСАНДРОВНА," "/1+)
ДАТА_(#20.02.1977,1977,ФЕВРАЛЬ,~20/2+)
Г_РОЖД(2-/3+)
ОБУЧ_("Образование:",ВЫСШИЙ/4+)
ОРГ_(ТАШКЕНТСКИЙ,ФИНАНСОВЫЙ,ИНСТИТУТ/5+)
"Образование:"(5-)
СПЕЦ_(ФИНАНСОВО,КРЕДИТНЫЙ,ФАКУЛЬТЕТ,СПЕЦИАЛЬНОСТЬ,ФИНАНСЫ,ПРЕДПРИЯТИЕ,
РАЗЛИЧНЫЙ,ФОРМА,СОБСТВЕННОСТЬ/6+)
ИМЕТЬ(5-,6-)
ВРЕМЯ_(1995,2000/7+)
ВРЕМЯ_РАБ(7-,5-)
ВРЕМЯ_(1994," "/8+)
ОРГ_(МАШИНОСТРОИТЕЛЬНЫЙ,ПРЕДПРИЯТИЕ,ПО,ИЗГОТОВЛЕНИЕ,ОБОРУДОВАНИЕ,
ТЕХИНПРОМ/9+)
"Профессиональный опыт"(9-)
PLACE_(РЕСПУБЛИКА,УЗБЕКИСТАН/10+)

ГДЕ(9-,10-)
ВРЕМЯ_(1996," "/11+)
ВРЕМЯ_РАБ(11-,9-)
РАБ_(БУХГАЛТЕР,КАССИР,МАТЕРИАЛЬНЫЙ,БУХГАЛТЕР,9-/12+)
.....
ПРЕДЛ_(0,РЕЗЮМЕ,Ф.,И.,О.,1-/42+) 42-(1,3,51)
ПРЕДЛ_(0,3-/43+) 43-(3,52,91)
ПРЕДЛ_(0,4-,5-,6-/44+) 44-(4,92,249)
ПРЕДЛ_(0,7-,8-,12-,НА,9-,10-/45+) 45-(6,250,484) .....

Фрагменты ДОК_(0,RESUME_1.TXT," "/0+) 0-(RUS) указывают, что содержательный портрет построен на основе рускоязычного текта из файла 'RESUME_1.TXT'. Следующие фрагменты представляет дату (ДАТА_), год рождения (Г_РОЖД), уровень квалификации (ОБУЧ_), организацию (ОРГ_) и ее свойство "Образование:", специальность (СПЕЦ_) и т.д. Фрагмент ИМЕТЬ(5-,6-) связывает учебное заведение и специальность, представленные в виде ОРГ_(.../5+) и СПЕЦ_(.../6+). Фрагменты ВРЕМЯ_(1995,2000/7+) и ВРЕМЯ_РАБ(7-,5-) представляют период времени учебы в упомянутом заведении. Особую роль играют фрагменты ПРЕДЛ_(...), которые соответствуют предложениям. Они заполняются словами, не вошедшими в информационные объектами, а также кодами самих объектов. К этим фрагментам добавляются указатели их местоположения в тексте. Например, фрагменты

ПРЕДЛ_(0,4-,"Образование:",5-,6-/44+) 44-(4,92,249)

представляют тот факт, что объекты с кодами 4+, 5+ и 6+ находятся в предложении, которое начинается с 4-ой строки текста и занимают место от 92-го до 249 байтов. Это средства позиционирования, которые необходимы для работы обратного ЛП. Итак, содержательные портреты - это наборы фрагментов РСС, с которые представляют достаточно высокий уровень формализации текстов и удобны для обработки - с помощью инструментальных средств ДЕКЛ [4].

3. Лингвистический процессор

Прямой лингвистический процессор (ЛП) системы LINGVO-MASTER обеспечивает автоматическое построение содержательных портретов. Он включает в себя лексикографический, морфологический, терминологический и синтактико-семантический анализ. Морфологический анализ необходим, чтобы избавиться от различных форм написания слов. Все словоформы одного и того же слова приводятся к единому виду - каноническому. Терминологический анализ обеспечивает выделение терминов, а также синонимичные преобразования. Синтактико-семантический анализ осуществляется специальными "контекстными" правилами, которые являются основой лингвистических знаний (ЛЗ). Контекстные правила позволяют выделять согласованные слова (словосочетания), а также несогласованные группы слов. Учитывается тот факт, что многие информационные объекты (даты, адреса, многие организации и т.д.) - это наборы слов, сокращений, мнемонических обозначений, которые часто грамматически никак не согласованы. Их выделение может осуществляться по чисто формальным принципам.

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 ФИО, глаголов и т.д. Выделение таких наборов слов (описаний объектов) основано на использовании контекстных правил следующего вида:

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

где <слово1>,... это может быть - отдельное слово, признак, а также И-ИЛИ графы. Для этих правил указывается, с какой позиции начинать применение, а также допустимый или недопустимый контекст. Далее, может быть указано, слово с какими признаками не должно стоять на той или другой позиции. Это обеспечивает дифференцированное применение правил. Такие правила выделяют из текста группы слов (по их признакам), описывающих какой-либо объект, и заменяют их на код фрагмента, например, представляющего адрес. Этот код рассматривается как самостоятельное слово со своими признаками. Правила применяются в определенной последовательности. Вначале выделяются объекты, затем их признаки, словосочетания, и наконец, глагольные формы. По мере применения таких правил строится семантическая сеть - содержательный портрет документа. Например, рассмотрим правило 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-) 3-(2,MORF) NOTBE(GG~1,2,LETT)

Это правило осуществляет преобразования:

ПРИЛАГАТЕЛЬНОЕ СУЩЕСТВИТЕЛЬНОЕ -> <комбинация слов> и

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). 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). Это пример наиболее простого правила. К таким правилам добавляются фрагменты, указывающие на контекст, на возможность каких-либо символов внутри и др. Специальные правила осуществляют идентификацию объектов, например, на основе местоимений или кратких описаний (по имени восстанавливается фамилия, если они где-нибудь упоминались вместе). И многое другое, что необходимо для работы с естественным языком.

Каждое контекстное правило - это семантическая сеть (РСС). Все лингвистические знания записываются в виде РСС. Над ними 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.

4. Обратный лингвистический процессор

Обратный ЛП служит для преобразования содержательных портретов (РСС) в компоненты ЕЯ и для их отображения на поля анкеты. Этот процессор имеет свои лингвистические знания (ЛЗ), с помощью которых задается последовательность выдачи рубрик (полей) и какими объектами они должны заполняться. Для выделения таких объектов служат их имена, а также связи, заданные в РСС. Приведем пример.

```
 W_SAY(ОРГ_," "," ","ПРОФЕССИОНАЛЬНЫЙ ОПЫТ",20/1+)
W_MUSTBE(1-,PROP," ","Профессиональный опыт")
1-(ВРЕМЯ_РАБ,1,ВРЕМЯ_,1,"Начало работы")
1-(ВРЕМЯ_РАБ,1,ВРЕМЯ_,2,"Окончание работы")
1-(" "," "," "," "," ","Название организации")
1-(РАБ_,OUTSIDE," "," ","Занимаемая должность") .....
```

Первые два фрагмента означают, что в содержательном портрете нужно искать ОРГ_(.../1+) со свойством "Профессиональный опыт"(1-). С помощью следующих двух фрагментов учитывается связь данной организации с временем работы. Они определяют поиск по коду ОРГ_(.../1+) другого фрагмента ВРЕМЯ_РАБ(2-,1-), где на 1-ом месте стоит ВРЕМЯ_(..,../2+). В нем надо взять первый аргумент и выдать под рубрикой "Начало работы", а второй - "Окончание работы". Далее под рубрикой "Название организации" выдается сама организация. Следующий шаг - это поиск фрагмента вида РАБ_(...,1-,..), включающего ОРГ_(.../1+). Он выдается под рубрикой "Занимаемая должность". Подобные ЛЗ задают стратегию "обхода" семантической сети по имеющимся связям с целью выделения объектов, которые выдаются под соответствующими рубриками (или заполняют поля анкеты).

Напомним, что информационные объекты содержат слова в канонической флрме. Выдавать их в таком виде для многих приложений недопустимо. Поэтому в системе LINGVO-MASTER выдача какого-либо объекта сводится к нахождению его месторасположения в тексте (см. п. 2) и поиску группы словоформ по каноническому представлению. Эта группа как бы вырывается из текста и помещается в поле анкеты. Таким образом, в полях анкеты будут только слова из текста. Это не относится к датам, которые преобразуются к единому виду, см. приложение 2. За счет ЛЗ обратного ЛП можно (сравнительно быстро) изменять поля и их содержимое, т.е. настраивать систему на определенную анкету или сайт. Конечно, содержимое полей формируется на основе информационных объектов, выделенных прямым ЛП. Если пользователю требуется поле с объектом, который не выделяется этим ЛП, то нужно дополнять анализ - вводить новые контекстные правила. В силу их независимости эта процедура также не является чрезмерно трудоемкой. В результате обеспечивается достаточно быстрая подстройка под область приложений. Помимо сказанного, LINGVO-MASTER содержит оболочку для построения экспертных систем. Экспертные знания вынесены в настройки. С их помощью анкета дополняется новыми данными: "Профессиональная область", ЯЗЫКИ (с указанием степени владения языком) и др., см. приложение 2.

В настоящее время система LINGVO-MASTER реализована в виде DLL-ки компанией "Новстрим" в рамках направления "Семантика - ИТ". Область ее приложений достаточно широка. Это прикладные программные системы, где требуется автоматическая формализация потока текстов на русском и английском языках. Особенности ее работы можно посмотреть, в частности, на сайте компании HEADHUNTER.RU - одного из ведущих рекрутинговых агентств в России.

Список литературы:

1. Кузнецов И.П. *Методы обработки сводок с выделением особенностей фигурантов и происшествий. Труды международного семинара Диалог-1999 по компьютерной лингвистике и ее приложениям. Том 2. Тарусса 1999.*

2. Кузнецов И.П. *Семантические представления. М. Наука. 1986г. 290 с.*

3. Kuznetsov Igor, Matskevich Andrey. *System for Extracting Semantic Information from Natural Language Text. Труды международного семинара Диалог-2002 по компьютерной лингвистике и ее приложениям. Том 2. Протвино, Наука, 2002.*

4. Кузнецов И.П. Пузанов В.В., Шарнин М.М. *Система обработки декларативных структур знаний ДЕКЛАР-2. Москва, ИПИАН, 1988 г.*

Приложение 1. Текст резюме:

Ф.И.О. Евгения Александровна Иванова.
Дата рождения: 20 февраля 1977 года
Образование высшее: Ташкентский Финансовый институт,
Финансово-кредитный факультет, специальность - <Финансы предприятий различных форм собственности>.
Время обучения: 1995-2000г.г.
Трудовая деятельность:
С 1994г- начала работать бухгалтером- кассиром, материальным бухгалтером на машиностроительном предприятии по изготовлению оборудования "Техинпром" - Республика Узбекистан.
С 1996г- переведена на должность главного бухгалтера на машиностроительном предприятии по изготовлению оборудования "Техинпром".
С 1999г. -2001г. - главный бухгалтер в компании "ВТФ Ипак Импекс" (первичная промышленная переработка табачного сырья)
Февраль 2002г. - июль 2002г. заместитель главного бухгалтера ООО<Промэлектроника М> - оптовая и розничная торговля (Москва)
Август 2002г. по наст.время - бухгалтер в LG Alina Electronics ltd., российский бух.учет (участок реализация, банк, касса, авансовые отчеты, книга покупок, книга продаж, расчеты с поставщиками и заказчиками, акты сверок), западный учет (банк - доходы, расходы)
Курсы повышения квалификации:
1993г.- прошла обучение в Учебно-научном центре <Корчалон> на <Отделении юристов школы предпринимателей>
1994г.- окончила курсы бухгалтеров в учебном центре <Алена-аудит>
Знание компьютера: "Platinum", "Accounting M-system" <1С:бухгалтерия> (версия 6.0), Microsoft Office (Microsoft Word, Microsoft Excel), Internet Explorer.
Знание языков: 1)английский язык - разговорный
2)корейский язык - чтение, письмо, перевод со словарем,
Водительское удостоверение категории <В>.
Контактный телефон в Москве: раб.903-61-91(доб.141)
Адрес электронной почты: zania_sheg@mail.ru

Приложение 2. Результат обработки:

Язык резюме:-> Русский
Фамилия:-> Иванова.
Имя:-> Евгения
Отчество:-> Александровна
Пол:-> Женский
Дата рождения:-> 20 февраль 1977
Возраст:->
Образование:-> 1 - (Высшее)
Желаемая должность:->
Профессиональная область:-> 5 - Банки-Инвестиции-Лизинг.
Специализация:->
Опыт работы:-> 2 - (3-6 лет)
Ожидаемый уровень месячного дохода:->
Регион:->
Возможность переезда:->
E-mail:-> zania_sheg@mail.ru
Почтовый адрес:->
Домашний телефон:->
Мобильный телефон:->
Рабочий телефон:-> 903-61-91(доб. 141)
Телефон:->
Личная интернет-страница:->
УЧЕБА:->
Название:-> Ташкентский Финансовый институт.
Факультет (специальность):-> Финансово кредитный факультет, специальность - <Финансы предприятий различных форм собственности>.

Диплом (степень):->
Начало учебы:-> 1995
Окончание учебы:-> 2000

ПРОФЕССИОНАЛЬНЫЙ ОПЫТ:->
Начало работы:-> 1996
Окончание работы:->
Название организации:-> Машиностроительном предприятии по изготовлению оборудования "Техинпром". Республика Узбекистан..
Занимаемая должность:-> Бухгалтером- кассиром, материальным бухгалтером.
Обязанность, функции, достижения:->

ПРОФЕССИОНАЛЬНЫЙ ОПЫТ:->
Начало работы:-> 1999
Окончание работы:-> 2001
Название организации:-> Ностроительном предприятии по изготовлению оборудования "Техинпром".
Занимаемая должность:-> Главного бухгалтера.
Обязанность, функции, достижения:->

ПРОФЕССИОНАЛЬНЫЙ ОПЫТ:->
Начало работы:-> Февраль 2002
Окончание работы:-> Июль 2002
Название организации:-> ООО Промэлектроника.
Занимаемая должность:-> Заместитель главного бухгалтера ООО Промэлектроника М> - оптовая и розничная торговля (Москва).
Обязанность, функции, достижения:->
ПРОФЕССИОНАЛЬНЫЙ ОПЫТ:->
Начало работы:-> Август 2002
Окончание работы:-> Настоящее время
Название организации:-> Alina Electronics ltd.
Занимаемая должность:-> Бухгалтер в LG.
Обязанность, функции, достижения:->
Российский бух. учет (участок реализация, банк, касса, авансовые отчеты, книга покупок, книга продаж, расчеты с поставщиками и заказчиками, акты сверок), западный учет (банк - доходы, расходы).
КУРСЫ (обучение):->
Проводящая организация:-> Учебно научном центре <Корчалон> на <Отделении юристов школы предпринимателей>.
Название курсов:->
Диплом (сертификат):->
Начало курсов:-> 1993
Окончание курсов:->
КУРСЫ (обучение):->
Проводящая организация:-> Учебном центре <Алена аудит>.
Название курсов:-> Курсы бухгалтеров.
Диплом (сертификат):->
Начало курсов:-> 1994
Окончание курсов:->

Тесты (экзамены):->
ЯЗЫКИ:-> ENGLISH - Английский, 3 - Могу проводить интервью
ЯЗЫКИ:-> 22 - Корейский, 2 - Читаю профессиональную литературу
РЕКОМЕНДАЦИИ:->
Ключевые навыки:->
Знание компьютера: `Platinum`, `Accounting M-system` <1С:бухгалтерия> (версия 6.0), Microsoft Office (Microsoft Word, Microsoft Excel), Internet Explorer.
Водительское удостоверение категории <B>.