Система выявления из документов значимой информации на основе лингвистических знаний в форме семантических сетей.

Кузнецов И.П., Кузнецов В.П., Мацкевич А.Г.

ИПИ РАН

igor-kuz@mtu-net.ru

Рассматривается система, ориентированная на обработку текстов естественного языка - сообщений средств массовой информации. Система выделяет из текстов семантически значимую информацию: интересующие пользователя объекты с их количественными и качественными характеристиками - атрибутами. Например, это могут быть производства с указанием их месторасположения, состава выпускаемой продукции, их количества, качества и т.д. Другие примеры: несчастные случаи с указанием причин (от травматизма, пожаров, дорожно-транспортных происшествий,...), количества погибших и характера повреждений; кономические показатели - номенклатура и количество выпускаемых изделий с указанием дат, места и др.

Выделяемые системой объекты и атрибуты определяются шаблонами, которые задаются пользователем. Каждый шаблон соответствует своему значимому объекту и состоит из связанных позиций (полей), которые сопоставляются атрибутам данного объекта. Каждый шаблон связан с лингвистическими знаниями, определяющими привязку его полей к компонентам естественного языка. Роль шаблона может играть таблица или схема базы данных (БД).

Задача системы - анализ данных ей текстов с заполнением полей введенных в нее шаблонов. Если роль шаблонов играют таблицы БД, тогда задача системы будет сводиться к автоматическому заполнению этих таблиц на основе данной ей текстовой информации.

Подобного сорта системы активно развиваются на Западе в рамках перспективных проектов, см. например, FASTUS [1]. Их перспективность определяется громадными объемами текстов, извлекаемых через ИНТЕРНЕТ, невозможностью для пользователя их прочитать или даже просмотреть в приемлемое время, чтобы найти интересующую его информацию.

Разработка такого сорта систем связана с серьезными трудностями, вызванными особенностями естественного языка, где присутствуют различные формы выражения одного и того же, многочисленные умолчания, анафорические ссылки и тд. Качественное выявление объектов с их атрибутами во многих случаях возможно только с широким привлечением семантики: тезаурусов, толковых словарей, средств синтактико-семантического анализа.

Обработка текста включает в себя:

- морфологический анализ;

- синтаксический анализ;

- контекстный анализ;

- семантический анализ и логико-аналитическая обработка.

Морфологический анализ имеет целью - приведение слов в каноническую форму.

Синтаксический анализ предложений с выделением словосочетаний и анализом форм осуществляется на основе специальных (продукционных) грамматик, записанных в форме семантических сетей [2]. Эта компонента поддерживается разработчиком системы.

Контекстный анализ осуществляется на основе слов-классификаторов и контекста [3]. По словам-классификаторам система распознает наличие объекта или его атрибута. Контекст определяет начало и конец атрибута (он может состоять из множества слов), а также знаки и слова, которые могут быть в соответствующих текстах описания.

Для контекстного анализа используются следующие лингвистические знания:

- типовые словосочетания;

- И-ИЛИ графы, задающие контекст;

- характеристические слова и ограничители.

В систему могут быть введены типовые словосочетания, которые рассматриваются как представляющие отдельный объект или атрибут, например, "Торговый дом", "Пенсионный фонд",... Такие словосочетания могут состоять из нескольких слов (не более 4-х).

При вводе такие словосочетания записываются в форме семантических сетей. Например, словосочетание "Торговый дом" представляется в виде фрагмента WORD("Торговый дом",ТОРГОВЫЙ,ДОМ).Такие фрагменты (за счет варьирования терминами в рамках их семантических пространств) позволяют отождествлять "Торговый дом" со словосочетаниями "Дом торговли", "Дом, в котором торгуют" и т.д. Если в БЗ имеется фрагмет SUB(ДОМ,ПАЛАТКА), указывающий, что ДОМ это может быть ПАЛАТКА, то "Торговый дом" будет выделяться из текстов, где говорится о торговых палатках.

И-ИЛИ графы задают фиксированные контексты значимых объектов. Например, конструкция ДАТА:<месяц><число><год или г.> указывает на составные части объекта ДАТА, которые могут встретиться в предложении. К такой конструкции добавляется указатель ее формы записи в виде семантических сетей. Объект ДАТА может содержать множество таких И-ИЛИ графов, задающих различные варианты контекста.

Характеристические слова позволяют находить части текста, описывающие значимый объект. Например, АДРЕС: АЛЛЕЯ, БУЛ.,ВАЛ, ДЕР., ЗДАНИЕ, КВ., КОР., КРАЙ и др.

Ограничители указывают на минимальное и максимальное количество слов, которые могут быть в предложении и описывать значимый объект. Помимо этого указываются возможные знаки пунктуации, формы записи слов-описателей, например, с большой буквы и др.

Далее осуществляется пост-лингвистическая и логико-аналитическая обработка, основанная на концепции семантических фильтров. В процессе такой обработки учитывается следующее.

Во-первых, слова-синонимы преобразуются к единому виду. Выделяются близкие по смыслу слова, например, которые приводят к одному результату. Словесное описание числовых

характеристик переводится в числовую форму.

Во-вторых, выделяются термины, а также словосочетания типа ВЫПУСК ТОВАРОВ, ПРИБЫЛЬ КОМЕРЧЕСКОГО БАНКА и др. Один и тот же признак или факт может быть выражен с помощью различных терминов и форм. Поэтому обработка включает элементы синтактико-семантического анализа. При этом учитываются близкие по смыслу слова, а также их ролевые функции и соотнесенность. Важно знать, что определяют слова типа ЗНАЧИТЕЛЬНЫЙ, НЕБОЛЬШОЙ,... – это прибыль или что-то другое.

В-третьих, различные люди по-разному называют одно и тоже. При выявлении значимых объектов системой обеспечивается варьирование терминами а рамках их семантических пространств. Система также учитывает, к примеру, что СВЕТЛЫЕ ВОЛОСЫ это могут быть РЫЖИЕ ВОЛОСЫ или БЛОНДИН и т.д. При этом используются элементы логического вывода. Например, по росту 190-195 см. система должна понимать, что человек ВЫСОГО РОСТА.

В-четвертых, важную роль играют, так называемые, качественные или содержательные признаки, например, отражающие вид аварии, причины потерь и др. Такие признаки могут в явном виде не присутствовать в текстах. Система обеспечивает их восстановление, используя соответствующие классификаторы, а также тезаурус, задающий семантические пространства терминов. Реализуется концепция управляемых семантических фильтров, которые срабатывают при наличии уточняющего материала. При этом осуществляется перебор многих вариантов, где допускаются перестановки слов текста, возможность их нахождения на определенном расстоянии. Служебные слова и предлоги (если они специально не заданы в фильтре) не учитываются. В результате охватывается большое количество примеров, которые могут встретиться в тексте.

Для семантического анализа используются терминологический словарь, представленный в предикатной форме, а также фрагментысемантической сети, управляющие работой семантических фильтров. Например, для выявления сумм затрат, убытка, долга, дохода,..., выраженных в различных валютах, используются фрагменты:

WORD("колич. денег",ЧИСЛО,ДЕНЬГИ,ЗАТРАТА)

SUB(ЗАТРАТА,УБЫТОК) SUB(ЗАТРАТА,ДОЛГ) SUB(ЗАТРАТА,ДОХОД)...

SUB(ВАЛЮТА) SUB(ВАЛЮТА,ДОЛЛАР) SUB(ВАЛЮТА,ФРАНК)...

Если (помимо сказанного) требуется выявление организаций с указанием их сумм затрат, убытка, то используются фрагменты:

WORD(ЗАТРАТА,ЗАТРАТА,ОРГАНИЗАЦИЯ)

SUB(ОРГАНИЗАЦИЯ,КООПЕРАТИВ) SUB(ОРГАНИЗАЦИЯ,ИЗДАТЕЛЬСТВО)...

Фрагменты типа WORD(...) - это обобщенные формы, допускающие различные вариации слов, представленных в родо-видовом дереве. Они могут быть различных типов: с указанием порядка слов или допускающие их свободный порядок. В последнем случае система будет выявлять слова, стоящие рядом на расстоянии в пределах 2-3-х позиций, что позволяет учесть разнообразные языковые формы с этими словами. Причем, вместо слов-признаков могут стоять их видовые понятия, пояснения, представленные фрагментами типа SUB.

Система нашла применение для автоматического построения баз знаний по сводкам происшествий, а также для выявления значимой информации, относящейся к коммерческой деятельности, по схеме: объекты (юридические лица, организации) - их атрибуты (числовые

показатели). Система в демо-варианте настроена на задачу выделения встречающихся в текстах числовых величин и категорий, к которым они относятся.

Литература

1. FASTUS: A Cacaded Finite-State Transducer for Extracting Information from Natural-Language Text. Rep. AIC, Menlo Park, California, 1996.
2. Кузнецов И.П. Семантические представления. М. Наука. 1986г. 290 с.
3. Кузнецов В.П., Мацкевич А.Г. Автоматическое выявление из документов значимой информации с помощью шаблонных слов и контекста. Труды международного семинара Диалог-98 по компьютерной лингвистики и ее приложениям. Том 2. Казань 1998.