УДК 519.95

Представление абстрактных объектов в базе знаний

Representation of abstract objects in the knowledge base

Курбатов С.С
Kurbatov S.S.

Анотация

В статье предлагается общая логика представления произвольных абстрактных объектов в базе знаний, формируемой в автоматизированном режиме в основном путем сканирования естественно-языковых текстов энциклопедического характера. Иерархия абстрактных объектов вычленяется автоматически путем выделения основного понятия в именной группе, описывающей вход в статью, и последующего поиска этого понятия как нового входа в статью энциклопедии. Выбраны специальные абстрактные объекты (математические формулы), для которых разработана логика их обработки и связи с материальными объектами. Автоматизированный режим предполагает ряд итераций, прерываемых анализом автоматически сформированных представлений и дополнением/модификацией эвристик, используемых при сканировании текстов.

Abstracts

In this paper the common logic of representation of any abstract objects in base of the knowledge formed in an automated mode in basic by scanning of the natural - language texts of encyclopaedic character is offered. The hierarchy of abstract objects is isolated automatically by allocation of the basic concept in nominal group describing an entrance in clause, and subsequent search of this concept as new entrance in clause of the encyclopedia. The special abstract objects (mathematical formulas) are chosen, for which the logic of their processing and connection with material objects is developed. The automated mode assumes a number of iterations, intrrupted by the analysis of the automatically generated representations and addition / updating heuristics, used at scanning texts.

Введение

В настоящее время центр тяжести в автоматической обработке естественного языка (ЕЯ) ощутимо сместился от задач машинного перевода и систем общения с ЭВМ (в частности ЕЯ-интерфейсы к базам данных) к задачам структуризации больших объемов текстовой информации. Практической целью такой структуризации является в основном высокоточный поиск информации в глобальной сети Интернета ([1], [2]), одна из наиболее продвинутых разработок - Fact Extractor [3]. Развиваемый в данной работе подход предполагает автоматизацию процесса формирования и сопровождения баз знаний по произвольным предметным областям путем сканирования соответствующих текстов на естественном языке.

Позиционирование подхода в общей схеме обработки текстов в системах принятия решений отражено на [4]. Развиваемый поход отличается от стандартного анализа, выполняемого системами структуризации большого корпуса текстов. Основное отличие заключается в том, что текст транслируется не в структуры базы данных, а в структуры базы знаний. Такая база знаний используется как для дальнейшей структуризации для целей, например, оперативного анализа (OLAP) и интеллектуального анализа (Data Mining), так и для диалога и получения отчетов непосредственно из базы знаний. Кроме того, в типичных структуризаторах достаточно развита техника обработки концептов типа "персона", "организация", "встреча" и т.п., однако техника для концептов типа "формула","переменная","уравнение" и т.п., развита в гораздо меньшей степени.

Предлагаемая логика ориентирована на автоматическое построение иерархии произвольных абстрактных объектов с возможностью последующего диалога на ограниченном естественном языке(ОЕЯ) по тематике этих объектов. Априорно задается минимальная иерархия, дающая самое общее (фундаментальное) разбиение абстрактных объектов на множества (не обязательно непересекающиеся). В процессе сканирования с помощью первоначальных базисных эвристик вычленяются

абстрактные объекты и их иерархия. Логика вычленения в целом аналогична общей логике автоматизированного построения иерахии и выделения материальных объектов [5]. В экспериментальном варианте выполнена программная реализаций предлагаемых решений.

Общая логика

Автоматическое построение иерархии предполагает (в первом приближении) продвижение по цепочке входов в статью энциклопедии до появления повтора входа и в этом случае повтор рассматривается как вероятное понятие высокого уровня. Примеры автоматически извлекаемых цепочек входов в статьи Большой Советской Энциклопедии (БСЭ), позволяющих в первом приближении выделить в качестве кандидатов на верхние уровни иерархии понятия “совокупность” (точнее множество) и “форма” (точнее содержание и форма).

А → буква → алфавит → совокупность = множество = Множеств теория = теория → комплекс → совокупность

А капелла → пение → искусство → форма → категория → значение → содержание → категории → понятие → форма

Отметим, что выделение понятий, входящих в цепочки, также выполнялось в автоматизированном режиме [6]. Разумеется, несколько уровней (морфология, синтаксис, материальные объекты и их характеристики) приводят к ряду ошибочных выводов программной системы. Однако развиваемый подход предполагает возможности автоматической коррекции ошибок, как путем модификации эвристик, так и путем сканирования дополнительного корпуса текстов (и лишь в минимальном объеме – исправление ошибок вручную).

Представление произвольных свойств абстрактных объектов сложнее, чем материальных. Поэтому в качестве первоочередных были выбраны специальные абстрактные объекты математической природы - формулы и последовательности, логика представления которых разработана более детально.

Априорная иерархия и специальные абстрактные объекты

Априорная иерархия в текущей версии включает: сознание, число, множество, последовательность, операция, формула, смысл, категория, материя, пространство, время, причина, следствие. Перечень априорных элементов носит предварительный характер и будет модифицироваться в процессе экспериментов. Эвристики верхнего уровня ссылаются на элементы априорной иерархии, чем обеспечивается выдвижение предположений после соотнесения некоторого конкретного объекта с элементом априорной иерархии. Примеры таких предположений, полученных экспериментально, приведены ниже.

Выбраны специальные абстрактные объекты, для которых разработана логика представления, позволяющая связывать эти объекты и фрагменты ЕЯ-описаний операций с этими объектами. В качестве первоочередных специальных абстрактных объектов выбраны формулы. Разработанное для них представление в базе знаний использует следующие приоритетные соображения:

- представление формулы в виде семантической сети;
- возможность представлять формулы, используя наиболее общие понятия (предикат равенства, переменные и функциональные символы, символы-сокращения для подформул и т.п.);
- возможность конкретизации входящих в формулу символов в процессе работы с системой;
- априорное задание базовых функциональных символов с соответствующей программной поддержкой, позволяющей проверять выполнимость формулы при конкретизации входящих в нее элементов;
- разбиение представления формул на структурную часть и декларативную часть (в первой задается собственно формула, во второй – описание входящих в нее элементов);
- ориентация на ЕЯ-обсуждение формулы (вопросы от тривиальных типа “Какие переменных входят в формулу?” до содержательных, например – “Каков физический смысл переменных формулы?”;
- возможность преобразования формулы в процессе ЕЯ-обсуждения;
- использование как графических, так и ЕЯ-описаний формул (например, $\int_a^b f(x)dx$ имеет ЕЯ-описание “интеграл от *a* до *b f* от *x dx*”) и возможности трансляции ЕЯ-описаний во внутренее семантическое представление.

Функциональные символы задаются либо с помощью композиции базовых (программно-поддержанных функций), либо таблично в виде строк значений переменных (возможно диапазонов) и результата. В последнем случае предполагается разработка механизмов для выдвижения (проверки) предположений о соответствии таблично заданной функции композиции базовых функций. Естественно, что эти механизмы аналогичны соответствующим общим механизмам по генерации предположений о морфологии, синтаксисе и построении иерархии материальных и абстрактных объектов.

Для входящих в формулу переменных предусмотрены возможности помимо описания наименования переменной, ее типа, позиции в

формуле и т.п. указывать физическую величину, соответствующую переменной. Этим обеспечивается связь абстрактного объекта “формула” с иерархией материальных объектов и, следовательно, с эвристиками, определяющими соответствующую “наивную” аксиоматику. Важно, что представление формулы в базе допускает большую степень неопределенности, которая, во-первых, не исключает ОЕЯ-обсуждения формулы, а во-вторых, позволяет в дальнейшем пополнить информацию о формуле и сделать обсуждение более содержательным.

Эксперимент

После получения иерархии абстрактных объектов в первом приближении работа проводилась в двух направлениях. Во-первых, разработка эвристик, позволяющих уточнить автоматически построенную иерархию (точнее, ее элементы). Во-вторых, использование построенных фрагментов иерахии для выдвижения предположений, связывающих объекты априорной иерархии с объектами ЕЯ-природы.

Отметим, что хотя тексты энциклопедических статей по сути ориентированы на иерархическую организацию, это тем не менее тексты реальной сложности. Трудности автоматической обработки были связаны как с ошибками при выделении ядра именной группы (вследствие ошибок при автоматическом определении морфологических характеристик и ошибок сегментации при определении входа в статью), так и ссылок вида “см. …”, “наименование …”, “один из …” и т.п.

Эксперимент включал как автоматическое построение иерархии абстрактных объектов, аналогично материальным (т.е. используя только вход в статью БСЭ и именную группу, предположительно описывающую вход), так и выявление кандидатов на формулы путем просмотра всех предложений БСЭ и выявление соответствующего контекста. Для последнего случая было выявлено более 400 контекстов, из которых с помощью эвристик были исключены варианты нематематических формул (химических, политических, ссылок и т.д.).

По оставшимся контекстам также с помощью эвристик была выявлена структурированная информация, записанная в базу знаний. Минимально записывалась ссылка на контекст и обоснование того, что это математическая формула (ссылка на соответствующую эвристику).

Пример автоматически выявленного контекста (приведен в угловых скобках):

< b, выраженное в секундах дуги, определяется формулой b = (206 264,8" u/c) sin g,

где u - скорость движения наблюдателя, c - скорость света и g - угол между направлениями на светило и апекс. >

Синтаксический анализ такого контекста, дополненный анализатором формул, позволяет выдвинуть предположения о виде формулы, входящих в нее переменных, их обозначении и физическом смысле и т.п. Детали эксперимента приведены в [ссылка на HTML-страницу]. Важно, что отсутствие полноценного описания (выявляемое автоматически) является для системы сигналом для пополнения описания формулы как путем диалога, так и путем поиска новых контекстов в других источниках (WORD-файлы с текстом и т.д.).

Уточнение эвристик

Эксперимент позволил выявить ряд недостатков в первоначально используемых эвристиках, которые устранялись соответствующей модификацией прежних эвристик или разработкой новых. Рутинные технические подробности таких уточнений приведены на HTML-странице [7]. Интересным идейным моментом было исправление ошибки в построенной иерархии ручным способом – исправление учителем (непосредственное указание на ошибку в режиме меню) с последующей попыткой программы обобщить исправление на другие элементы иерархии.

Помимо формул была разработана примитивная логика представления и работы с таким фундаментальным понятием как *последовательность*. Стиль такого представления ориентирован на использование механизмов выдвижения предположений о морфологии и синтаксисе в более общих ситуациях. Например, после разбора входа и первой именной группы БСЭ предложения “а – первая буква русского алфавита” были разработаны эвристики, позволяющие выдвигать предположения для остальных букв. При этом использовалось общее представление алфавита как *последовательности* букв и общих методов, применимых к последовательности – “выдать номер заданного элемента” и “выдать элемент по номеру. Развитие инструментальных средств, обеспечивающих исследователю комфортную среду для проведения экспериментов, намечено в [8].

Заключение

Данная работа выполнена в рамках общего подхода, ориентированного на создание инструментальных программных средств, позволяющих автоматизировать извлечение знаний собственно о языке, знаний общего характера о внешнем мире и конкретных знаний в данной предметной области. Такие инструментальные средства должны сканировать произвольные ЕЯ-тексты, связывая извлекаемые из них объекты с элементами фундаментальной (априорно заданной) иерархии. Наследование наивной аксиоматики, определяемой в данной иерархии, позволит выдвигать предположения об извлекаемых из текстов объектах и подтвер-

ждать/опровергать эти предположения также путем сканирования текстов.

Количественные характеристики (объем) априорно задаваемой иерархии будет уточняться в процессе эксперимента с учетом прикладных аспектов развиваемого подхода. Практическое использование результатов работы предполагает выявление ошибок и неточностей в текстах с большим содержанием произвольного вида формул (диссертации, научные статьи, отчеты, технические задания и т.п.). Спектр анализа весьма широк – от выявления несоответствий единиц измерения до числовых ошибок.

Литература:

1. *Липинский Ю.В.* , Средства информационного поиска и навигации в больших массивах неструктурированной информации, компания “ Гарант-Парк-Интернет ”.
2. *Нечипоренко А.В., Русин А.О.* Система автоматизированного извлечения знаний из текстов на естественном языке // Труды международной научно-технической конференции. “Информационные системы и технологии – 2003”, г. Новосибирск. ", Компания «НооЛаб». 12.08.2008.
3. *Михайленко П.* Язык онтологий в Web // Журнал "Открытые системы". 2004 г. №2
4. HTML-страница, URL: http://eia--dostup.ru/shema_01.htm
5. *Курбатов С.С.*, Априорная модель данных в реляционных базах // Новости искусственного интеллекта № 6, М.: Анахарсис, 2003.
6. *Курбатов С.С.*, Автоматизированное построение естественно-языкового интерфейса для реляционных баз данных // Новости искусственного интеллекта № 2, М.: Анахарсис, 2002, С. 17-21.
7. HTML-страница, URL: http://eia--dostup.ru/head_doc_01.HTM
8. *Курбатов С.С.* Инструментальные средства для автоматизированного формирования баз знаний // Труды 5-ой международной научно-практической конференции "Интегрированные модели и мягкие вычисления в искусственном интеллекте", Коломна. М.: 2009.