Интеграционный подход к автоматизированному формированию лингвистических баз знаний

М.Г. Мальковский, А.В. Субботин, Т.Ю. Грацианова

Факультет вычислительной математики и кибернетики МГУ им. М.В.Ломоносова, кафедра алгоритмических языков
119899, Москва, ГСП, Воробьевы горы, МГУ, факультет ВМиК
malk@cs.msu.su, Subbotin@isd.srcc.msu.su

На сегодняшний день очевидна необходимость построения и внедрения систем, обрабатывающих естественный язык (ЕЯ-систем). Такие системы применяются практически во всех областях человеческой деятельности, и решают широкий спектр задач, начиная от поддержки создания текстов, до обработки запросов на естественном языке (ЕЯ) и распознавания и синтеза речи [Cole-1995]. Общепризнанным фактом является то, что для достижения приемлемого для большинства задач качества ЕЯ-обработки необходимы специальные информационные массивы, содержащие информацию о языке - так называемые лингвистические базы знаний (ЛБЗ).

Формирование ЛБЗ представляет собой процесс сбора лингвистической информации, представления ее в виде, пригодном для автоматической обработки и поддержание этой информации в актуальном состоянии. Сбор лингвистической информации осуществляется на основе источников, которые могут быть классифицированы следующим образом:

- Люди
  - Эксперты-лингвисты
  - Носители языка
- Тексты
  - Описания языка, созданные специалистами
  - Тексты, не являющиеся описаниями языка.

Использование всех этих источников является необходимым, поскольку ни один из них полностью не дает в отдельности информации обо всем спектре языковых явлений [Хейс-1971]. Эксперты-лингвисты уделяют пристальное внимание в основном весьма специфичным явлениям языка. Обычные носители языка редко способны формулировать достаточно четко свои знания о языке. Создание описаний специалистами – длительный и очень трудоемкий процесс, и эти описания часто "не успевают" за развитием языка, поэтому приходится извлекать информацию также и из обычных текстов.

Из вышесказанного следует, что формирование ЛБЗ представляет собой специальный вид ЕЯ-обработки и является достаточно сложным и трудоемким процессом, требующим как выполнения множества рутинных операций, так и привлечения знаний и навыков экспертов. В связи с этим актуальной является проблема автоматизированного[1] формирования ЛБЗ.

---

[1] Под "автоматизированным" формированием ЛБЗ будем понимать процесс формирования ЛБЗ, выполняемый совместно человеком и машиной - в отличие от "автоматического" - предусматривающего, что задача может быть решена без участия человека.

Практика последних лет показала, что подходы к ЕЯ-обработке, основанные на поверхностной информации о языке, например статистической, способны обеспечить лишь ограниченный уровень качества обработки (количества правильно распознанных слов, выданных релевантных документов и т.п.). Для повышения качества обработки требуется применение более сложной и лингвистически содержательной информации [Мальковский-1996, Church-1995].

В последнее время четко обозначилась тенденция к повышению качества ЕЯ-обработки за счет совместного применения различных (традиционных и новых) методов. Такой подход широко применяется и в нашей исследовательской группе. Дают наибольший эффект и в то же время представляют наибольшую трудность для применения методы, основывающиеся на комбинациях различных, часто далеких друг от друга подходов. Так, например, комбинация статистических методов и методов локального синтаксического анализа позволяет обеспечить точность обработки свыше 95% в системах распознавания речи [Malkovsky-1997]. Аналогичные тенденции, однако, гораздо более ярко выраженные, наблюдаются и в области автоматизированного формирования ЛБЗ. На сегодняшний день интеграция различных методов становится необходимым условием успешного решения этой задачи [Chen-1994, Gref-1993, Cardie-1997].

Среди основных факторов, осложняющих интеграцию различных методов обработки, отметим наиболее существенные:

- Реализация различных методов формирования ЛБЗ осуществляется на базе различных информационных технологий, часто без опоры на стандарты.
- Отсутствуют целостные подходы к интеграции различных методов, соответствующие лингвистические и математические модели (представления данных и процессов обработки).

В этой работе описывается попытка обобщения опыта интеграции различных методов автоматизированного формирования ЛБЗ и разработки целостного подхода. Для этого формулируется метамодель, на базе которой могут быть интегрированы различные методы обработки текстов. Описываются также основные принципы построения системы автоматизированного формирования ЛБЗ, реализующей предлагаемый подход.

Основы предлагаемого подхода

Рассмотрим указанные проблемы интеграции методов автоматизированного формирования ЛБЗ и пути их решения на примере построения тезауруса.

Различные методы формирования используют различные модели представления данных. Так, например, при автоматизированном формировании тезауруса информация о текстовых связях терминов может представляться в виде матриц совместной встречаемости [Chen-1994]. Традиционный же тезаурус обычно представляется в виде совокупности словарных статей - заглавной тезаурусной единицы и связанных с ней других тезаурусных единиц. Альтернативным представлением тезауруса является представление тезауруса в виде графа. На самом деле эта информация описывает одни и те же или похожие явления, однако, представляется она в различных видах.

Процесс извлечения тезаурусных отношений также может представляться различным образом. Обычно при описании каждого из методов

дается алгоритм. Такого рода алгоритмы часто очень похожи по своей организации: можно выделить несколько одинаковых этапов, отличающихся в основном применяемыми эвристиками. Однако из-за того, что эти эвристики жестко связаны с алгоритмом, использовать их отдельно от алгоритма достаточно сложно. С другой стороны, именно комбинации этих эвристик, правил формирования, предлагаемых в рамках различных методов, дают обычно наиболее высокие результаты. Причем, достоверно установить, какая комбинация даст наилучший результат можно только экспериментальным путем.

Следовательно, одним из подходов к интеграции методов автоматизированного формирования ЛБЗ может быть следующий:

- выработка унифицированной модели ЛБЗ, или метамодели (представления данных и процесса формирования);
- отображение существующих методов на построенную метамодель;
- выбор оптимальной комбинации методов путем экспериментов на основе выбранной модели.

Реализуются данные методы обычно на различных технологических платформах: языках программирования, СУБД, программном обеспечении промежуточного слоя. В связи с этим, помимо отображения различных методов на общую модель, надо решить проблему совместного использования различных массивов данных, алгоритмов, написанных на различных языках и т.д. Для решения этой проблемы может быть проделан следующий путь:

- создание технологии описания ЛБЗ и процесса ее формирования, а также технологии интеграции методов обработки;
- реализация инструментальной системы, поддерживающей данную технологию;
- интеграция компонентов, реализующих различные методы в данную систему.

Следует отметить, что технология описания и система должны базироваться на стандартах, в противном случае крайне сложно будет использовать данную систему в условиях постоянного развития технологий создания программного обеспечения.

Можно сформулировать следующие требования к метамодели:

- Представление неточной информации. Это требование обусловлено тем, что большая часть лингвистической информации, особенно относящейся к семантическому и синтаксическому уровням, не является точной, а носит статистический или эвристический характер.
- Достаточная мощность для представления различных моделей данных и процессов обработки.

На основе требований к модели сформулируем основные требования к системе автоматизированного формирования ЛБЗ, реализующей данную модель:

- Особенности источников данных (например, форматы хранения, интерфейсы доступа и т.п.) должны быть инкапсулированы таким образом, чтобы можно было подключать новые и заменять существующие источники данных путем замены отдельных компонентов системы.
- Система должна поддерживать интеграцию различных обрабатывающих компонентов и источников данных и настройку без перекомпиляции и перепрограммирования.

Формирование ЛБЗ требует обработки большого количества данных. Принимая во внимание, что при унификации различных моделей и процессов обработки часто придется жертвовать эффективностью, необходимо обеспечить высокую масштабируемость. Учитывая это и сформулированные требования, можно сделать вывод о том, что система должна поддерживать распределенную обработку.

Метамодель

Информация в предлагаемой метамодели представляется в виде семантической сети, в узлах которой находятся структуры, схожие с фреймами Минского [Minsky-1974]. Сходное представление информации используется в классических методологиях современного объектного подхода. Такой подход к представлению знаний является также традиционным для искусственного интеллекта и хорошо себя зарекомендовал в современной компьютерной лингвистике.

Процесс формирования ЛБЗ в метамодели представляется в виде схожем с процессом логического вывода, также широко используемого при решении задач искусственного интеллекта. Действительно, большинство исследуемых в компьютерной лингвистике типов объектов является общепризнанным и достаточно статичным. Такие лингвистические объекты как лексемы, семы, словоформы, предложения, фигурируют практически во всех работах и их определения отличаются в основном деталями. Новая, динамичная и трудно извлекаемая информация в основном описывает связи между этими объектами. Связи между объектами могут быть представлены отношениями (в рамках предлагаемой метамодели) или предикатами. Таким образом, процесс формирования ЛБЗ сводится в основном к поиску отношений между объектами в источниках данных, получению информации о новых отношениях путем анализа существующих отношений и отслеживанию их изменений.

Например, в случае автоматизированного формирования тезауруса, информация о семантических связях понятий устанавливается путем анализа текстовых связей (в обычных текстах), связей терминов в толковых словарях и связей между понятиями и выражающими их лексическими единицами. На базе этого анализа по некоторым правилам делаются заключения о наличии семантических отношений между понятиями.

Предлагаемая модель основывается на следующих аппаратах:

- Язык UML [OMG ad/99-02-01] - фактически стандартный язык объектного моделирования, вобравший в себя наиболее прогрессивные идеи и средства моделирования, объектного и структурно-функционального подходов. К тому же этот язык имеет строго описанную семантику и мощную объектную модель. Одним из интереснейших с точки зрения решения поставленной задачи свойств UML является его расширяемость. В описании языка предусмотрены средства введения новых средств моделирования. Указанные свойства делают его хорошим базовым аппаратом для построения метамодели ЛБЗ.
- Нечеткая логика [Тэрано-1993]. С помощью идей, основанных на нечеткой логике (и других разделах нечеткой математики) можно компактно представлять как статистическую информацию, так и информацию, имеющую неточный характер (например, мнения экспертов). Процесс

нечеткого вывода (прямого или обратного) позволяет описывать преобразования нечеткой информации.

Метамодель строится следующим образом. Выделяется подмножество метамодели UML, необходимое для моделирования ЛБЗ. Это подмножество адаптируется для интеграции с формализмом нечеткой математики с помощью средств расширения языка: введения новых стереотипов и связанных с ними ограничений и свойств (tagged values). С помощью модифицированной метамодели создается каркас моделей ЛБЗ (framework), включающий в себя классы и другие элементы моделей, на базе которых с помощью наследования (специализации) будут строиться описания ЛБЗ. На основе этой метамодели с использованием элементов моделей, входящих в framework, осуществляется моделирование конкретных ЛБЗ.

Далее коротко рассмотрим используемое подмножество языка UML и то, каким образом оно используется.

Диаграммы статической структуры (Static Structure Diagrams) используются для описания модели данных ЛБЗ. Для ассоциаций UML вводится специальный стереотип, позволяющий описывать нечеткие отношения (например, задавать их формулой, выражая через другие нечеткие отношения, или указывая их вес константой и т.п.). Для классов ассоциаций вводится базовый класс, в частности позволяющий указывать значение функции принадлежности каждого экземпляра отношения.

С помощью диаграмм Use Case (диаграмм сценариев использования) представляются процессы (сценарии) формирования ЛБЗ на верхнем уровне абстракции. Каждый из этих процессов либо выполняется целиком, либо не выполняется вообще. Процессы формирования ЛБЗ инициируются внешними по отношению к системе формирования ЛБЗ сущностями - экспертами или другими системами.

С каждым Use Case связывается диаграмма действий (Activity Diagram), с помощью которой раскрывается процесс формирования ЛБЗ. Состояния этой диаграммы могут описывать либо шаги нечеткого вывода, либо некоторые другие, внешние по отношению к выводу действия (ввод, вывод, преобразования информации). С каждым состоянием может быть связана процедура, осуществляющая необходимые вычисления. Таким образом, процесс формирования ЛБЗ представляется в виде процесса нечеткого вывода, расширенного некоторыми действиями. С помощью переходов указываются зависимости, позволяющие определить порядок анализа информации и вычисления отношений.

Рассмотрим, как данная метамодель используется для моделирования процесса автоматизированного формирования тезауруса. Объекты тезауруса одинаковы практически во всех методах формирования и моделируются классами. В основном методы отличаются набором анализируемых текстовых связей и правилами получения по ним тезаурусных связей. Для каждого из методов может быть описан набор связей и один или несколько Use Case, представляющих процесс формирования ЛБЗ. При описании процесса формирования с состояниями связываются процедуры вычисления различных текстовых связей, реализованные в рамках различных методик. Это могут быть связи по совместной встречаемости в "окне" того или иного размера, синтагматические связи и т.п.

Будучи реализована, такая модель позволяет легко создавать новые процессы формирования ЛБЗ и проводить эксперименты, комбинируя различные методы, путем операций над отношениями и их зависимостями на уровне UML моделей.

Система автоматизированного формирования ЛБЗ

Система автоматизированного формирования ЛБЗ создается на базе стандарта CORBA [OMG 96.03.04]. Метамодель представления данных реализуется настраиваемым и расширяемым CORBA сервером (написанным на C++). Этот сервер позволяет декларативно определять схему данных на основе сформулированной метамодели и осуществлять ее отображение на модель источника данных (например, на схему реляционной СУБД). В этот сервер могут интегрироваться компоненты, обеспечивающие работу как с различными объектами и отношениями, так и специальные компоненты, реализующие отдельные процедуры формирования ЛБЗ.

Метамодель процесса обработки реализуется совокупностью CORBA серверов, обеспечивающих координацию работы различных обрабатывающих компонентов (которые должны поддерживать специальный IDL интерфейс) в рамках концепции workflow. С помощью этого сервера в частности реализуется процесс нечеткого вывода.

Основные этапы технология настройки данной системы на процесс формирования ЛБЗ следующий:

1. На языке UML строятся модели в рамках описанной выше метамодели.
2. С помощью UML CASE средства осуществляется генерация описаний данных и процессов, используемых системой. Также может генерироваться код обрабатывающих компонентов.
3. Данные описания закладываются в систему с помощью специальных механизмов настройки. При этом в систему интегрируются требуемые источники данных и обрабатывающие компоненты.

После выполнения этих процедур система готова к работе (или экспериментам).

Литература

[Мальковский-1996] Мальковский М.Г., Субботин А.В. Синтаксический анализ в прикладных ЕЯ системах // "Интеллект. Язык. Компьютер", вып.4, Казань, 1996. - С. 67 - 72.

[Тэрано-1993] Прикладные нечеткие системы. Под ред. Тэрано Т., Асаи К., Сугено М., М.: Мир, 1993.

[Хейс-1971] Хейс Д.Г., Методы исследования в области автоматического перевода // Автоматический перевод, М.:1971, С. 41-83.

[Cardie-1997] Cardie, S. Proposal for a Framework for the High-Precision Identification of Linguistic Relationships, 1997.

[Chen-1994] Chen,H., Schatz,B., Yim,T., Fye,D. Automatic Thesaurus Generation for Electronic Community System, University of Arizona, USA, 1994.

[Church-1995] Church, K. And Rau, L. (1995) Commercial Applications of Natural Language Processing, in CACM, Vol. 38, No.11, 1995

[Cole-1995] Cole,R., Mariani,J., Uszkoreit, H., Zaenen, A.,Zue, V. Eds (1995) Survey of the State of the Art in Human Language Technology ftp://speech.cse.ogi.edu/pub/docs/HLT.

[Gref-1993] Grefenstette,G., Hearst,M. A Method for Refining Automatically-Discovered Lexical Relations: Combining Weak Techniques for Stronger Results.

[Malkovsky-1997] Malkovsky M. NL-Processor in a Speech Recognition System // On-line Conference "Speech Synthesis and Analysis" - 1997 - http://www.kcn.ru/tat_en/science/fccl

[Minsky-1974] Minsky,M. A Framework for Representing Knowledge, Massachusetts Institute of Technology, 1974

[OMG 96.03.04] The Common Object Request Broker Architecture and Specification. Revision 2.0. - OMG Document 96.03.04, July, 1995.

[OMG ad/99-02-01]: UML 1.3 alpha R2 draft – OMG Document ad/99-02-01, February, 1999.

TOWARD INTEGRATED APPROACH TO COMPUTER-AIDED BUILDING OF THE NATURAL LANGUAGE KNOWLEDGE BASES

Malkovsky,M., Subbotin,A., Gratsianova,T.

A Natural Language Knowledge Base (that describes the items and rules of lexical, syntactic and semantic levels) is considered to be an important part of any text/speech processing system for real applications. Some aspects of the problem of computer-aided building and maintenance of the Natural Language Knowledge Bases are discussed. The main problem of the state of the art NLKB building is the integration of different methods. In this paper we describe an approach to such integration. The main component of this approach is UML and fuzzy mathematics based metamodel of the knowledge acquisition process and the information representation. Requirements to the software implementation are described too.