

A 05.13.02.
M-241

ՀԱՅԱՍՏԱՆԻ ՀԱՆՐԱՊԵՏՈՒԹՅԱՆ ԿՐԹՈՒԹՅԱՆ ԵՎ ԳԻՏՈՒԹՅԱՆ
ՆԱԽԱՐԱՐՈՒԹՅՈՒՆ

ՀԱՅԱՍՏԱՆԻ ՊԵՏԱԿԱՆ ՃԱՐՏԱՐԱՊԻՏԱԿԱՆ ՀԱՄԱԼՍԱՐԱՆ

ՄԱՆՈՒԿՅԱՆ ԱՐԱՄ ՍՈՒՐԵՆԻ

ԲԱՐՁՐ ՇԱՀՎՈՐԾՈՂԱԿԱՆ ԲՆՈՒԹԱԳՐԵՐ ԱՊԱՀՈՎՈՂ ՄԵՔԵՆԱՅԱԿԱՆ
ԹԱՐԳՄԱՆՈՒԹՅԱՆ ԿԱԶՄԱԿԵՐՊՍԱՆ ՄԵԹՈՂՆԵՐ

Ե.13.02 - «Ավտոմատացման համակարգեր»
մասնագիտությամբ տեխնիկական գիտությունների թեկնածուի գիտական
աստիճանի հայցման ատենախոսության

ՍԵՂՄԱԳԻՐ

Երևան 2006

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РЕСПУБЛИКИ АРМЕНИЯ

ГОСУДАРСТВЕННЫЙ ИНЖЕНЕРНЫЙ УНИВЕРСИТЕТ АРМЕНИИ

МАНУКЯН АРАМ СУРЕНОВИЧ

МЕТОДЫ ОРГАНИЗАЦИИ МАШИННОГО ПЕРЕВОДА ТЕКСТОВ С ВЫСОКИМ
УРОВНЕМ ЭКСПЛУАТАЦИОННЫХ ХАРАКТЕРИСТИК

АВТОРЕФЕРАТ

диссертации на соискание ученой степени кандидата технических
наук по специальности

05.13.02 – «Автоматизированные системы»

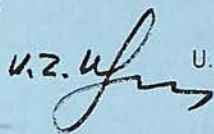
Ереван 2006

Ատենախոսության թեման հաստատվել է Հայաստանի պետական ճարտարագիտական համալսարանում

Գիտական ղեկավար՝ տ.գ.դ., պրոֆեսոր Է.Ն. Մանուկյան

Պաշտոնական ընդդիմախոսներ՝ ֆ-մ.գ.դ., պրոֆեսոր Շ.Ե.Բոգոյան
ֆ-մ.գ.թ Վ.Գ. Սահակյան

Առաջատար կազմակերպություն՝ «Եվրոպական կրթական տարածաշրջանային ակադեմիա»

Պաշտպանությունը կայանալու է 2006թ. հունիսի 30-ին ժամը 14⁰⁰ -ին Հայաստանի պետական ճարտարագիտական համալսարանի 032 Սասնագիտական խորհրդի նիստում: Հասցեն՝ 375009, Երևան, Տերյան փ.105: Ատենախոսությանը կարելի է ծանոթանալ ՀՊԵՀ-ի գրադարանում: Սեղմագիրն առաքված է 2006թ. մայիսի 30-ին: Սասնագիտական խորհրդի գիտական քարտուղար, տ.գ.դ., պրոֆեսոր՝  Ս.Ր.Սիմոնյան

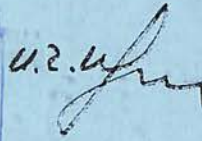
Тема диссертации утверждена в Государственном инженерном университете Армении

Научный руководитель: д.т.н., проф. Э.Н. Манукян

Официальные оппоненты: д.ф-м.н., проф. Ш.Е.Бозоян
к.ф-м.н. В.Г.Саакян

Ведущая организация: “Европейская региональная академия”

Защита состоится 30-го июня 2006г. в 14⁰⁰ часов на заседании Специализированного совета 032 Государственного инженерного университета Армении по адресу: 375009, Ереван, ул. Теряна 105. С диссертацией можно ознакомиться в библиотеке ГИУА. Автореферат разослан 30-го мая 2006г.

Ученый секретарь Специализированного совета, д.т.н., проф.  С.О. Симонян



2434-2006

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность работы. Компьютерное моделирование языковой деятельности человека является одной из базовых проблем в области построения интеллектуальных систем. Необходимость в подобной деятельности возникает:

- в задачах компьютерной обработки больших массивов естественно-языковых текстов (ЕЯ-текстов);
- при организации естественно-языкового интерфейса компьютерных систем с пользователем;
- в процессе создания больших банков информации на основе естественных языков;
- для разработки языков - посредников в многоязычной информационной среде.

Основным компонентом средств решения подобных задач является система машинного перевода (СМП). Эффективность функционирования СМП определяется возможностями ее математического обеспечения. Реальные же возможности для дальнейшего развития средств математического обеспечения СМП с целью улучшения ее эксплуатационных характеристик и повышения функциональных возможностей и доведения ее до уровня систем, оснащенных искусственным интеллектом, достигаются благодаря наблюдающемуся в последние годы качественному улучшению технических характеристик ЭВМ.

Целью данной работы является разработка методов создания автоматизированных систем машинного перевода, имеющих высокий уровень адаптации внутримашинных процессов организации процесса перевода к изменениям внешней среды функционирования СМП и обеспечивающих высокий уровень их эксплуатационных характеристик.

Работы по созданию программного обеспечения переводчика велись компанией ISMA в течение 2000-2003 гг., а лингвистическая настройка и формирование его базы знаний продолжались в лаборатории машинного перевода ГИУА с 2004 года. В результате была создана база знаний СМП, содержащая порядка 35000 концептов и примерно 50000 словарных статей (имен концептов или их синонимов). В процессе практических разработок СМП выявились задачи, от решения которых в наибольшей мере зависит эффективность ее функционирования. Решению некоторых из них, наименее освещенных в литературе, и посвящена диссертационная работа.

Задачами исследования являются:

1. Разработка схемы функционирования СМП, обеспечивающей полную сохранность смысла входного предложения в выходном тексте и независимость грамматики входного текста от выходного.
2. Разработка "послойной" архитектуры программных средств, обеспечивающей:
 - максимальную независимость слоев, что позволит организовать технологическую независимость процессов модификации программных средств СМП в течение их развития;
 - возможность для подключения к системе новых отладочных средств, ориентированных на пользователя-лингвиста;
 - соблюдение принципа синтаксической управляемости программных модулей ядра переводчика (в качестве управляющей компоненты может выступать непроцедурное описание входного и выходного языков).
3. Описание синтаксиса языка, максимально ориентированного на организацию двух основных этапов перевода: синтаксически-семантического анализа и конечного семантического анализа.
4. Разработка логической организации базы знаний (БЗ).

5. Разработка физической организации базы знаний и базы данных системы, обеспечивающих близкое к минимуму время доступа к данным.
6. Разработка статистического метода контекстного анализа, способствующего в определенной степени решению проблемы омонимии.
7. Разработка методов процесса генерации семантического аналога входного предложения для последовательности концептов (при заданных ограничениях на временные ресурсы), исключающих полный перебор и обеспечивающих максимальную вероятность правдоподобности выбранной последовательности.
8. Разработка метода и алгоритма определения наикратчайших путей в семантической сети базы знаний для организации конечного семантического анализа.
9. Разработка метаязыка, ориентированного на организацию анализа текстов с контекстно-зависимой грамматикой.

Научная новизна:

1. Разработана схема организации поверхностно-вероятностного анализа, позволяющая автоматически установить тематику анализируемых фрагментов текста и, тем самым, повысить эффективность решения задачи омонимии. Предложенное содержит признак новизны, поскольку в большинстве распространенных переводчиках тематика устанавливается вручную. Получены теоретические оценки потерь машинного времени для известной функции распределения достоверности концептов, претендующих на выбранное слово. На этапе эскизного проектирования полученные оценки позволяют проектировщику СМП выбрать оптимальную глубину поверхностно-вероятностного анализа. При этом обоснованно минимизируются потери на данном этапе и уменьшаются ожидаемые потери на последующем этапе детерминированного анализа последовательности концептов (ПК).

2. Впервые предложен метод формирования плана генерации вариантов ПК частичного перебора. При заданных ограничениях на временные ресурсы метод обеспечивает минимальную вероятность пропуска наиболее достоверного варианта ПК.
3. Предложены метод и алгоритм определения наикратчайших путей в семантической сети базы знаний для организации конечного семантического анализа.
4. Разработан новый метаязык, ориентированный на организацию анализа текстов с контекстно-зависимой грамматикой.

Практическая значимость работы. Данная работа выполнялась в рамках госбюджетных и проектных работ ГИУА, в том числе:

1. Госбюджетная тема N0294 “Автоматизированная система машинного перевода”.
2. Грант № 13290, Y03-9006 по теме “Создание программного обеспечения электронного армяно-английского и англо-армянского переводчика”.

Приведенные в работе методы и алгоритмы могут быть использованы:

- при создании СМП;
- в системах речевого управления в робототехнике;
- при построении подсистем анализа текстов на ЕЯ в различных прикладных системах;
- в учебных дисциплинах “Системы искусственного интеллекта”.

Реализация результатов работы. Созданные схемы организации работ программных средств, методы и алгоритмы были использованы при разработке двухстороннего армяно-английского переводчика, который функционирует по схеме: “предложение входного языка” -> “семантический граф, представляющий смысл входного предложения” -> “предложение на выходном языке”. В процессе

создания переводчика были разработаны синтаксически управляемые N-конвертор, D-конвертор и база знаний СМП, содержащая примерно 35000 концептов и порядка 50000 словарных статей (имен концептов или их синонимов). Переводчик установлен в Интернете (Web адрес: <http://www.translator.am>) и работает в on-line режиме.

Апробация работы. Научные результаты исследований докладывались и обсуждались на:

- Конференции CSIT-2003. Ереван, Армения.
- Годичной научной конференции ГИУА (2001, 2002, 2004).
- Международной молодежной конференции (2005).

Публикации. Основные научные результаты работы опубликованы в девяти научных трудах, список которых представлен в конце автореферата.

Структура и объем диссертационной работы. Работа состоит из введения, пяти глав, заключения, списка литературы из 101 наименования. Работа включает 27 рисунков. Общий объем работы - 115 страниц.

На защиту выносятся следующие положения:

1. Схема организации поверхностно-вероятностного анализа системы машинного перевода.
2. Метод формирования плана генерации вариантов последовательности концептов, обеспечивающий минимальную вероятность пропуска наиболее достоверного варианта последовательности концептов при заданных ограничениях на временные ресурсы.
3. Метод и алгоритм определения наикратчайших путей в семантической сети базы знаний для организации конечного семантического анализа.
4. Метаязык, ориентированный на организацию анализа текстов с контекстно-зависимой грамматикой.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

В первой главе выполнен обзор известных систем машинного перевода, обсуждены их технические и эксплуатационные характеристики. Рассматриваются методы анализа текстов на ЕЯ, методы организации N-конверторов и D-конверторов, а также базы знаний, являющиеся основным компонентом современных СМП. На основе анализа известных методов организации работ их основных компонентов, условий их разработки, внедрения и эксплуатации СМП выработан комплекс требований, предъявляемых к функциональным возможностям и эксплуатационным характеристикам средств математического обеспечения основных компонентов СМП, с учетом достижений технических характеристик современных компьютеров. Сформированные задачи, вытекающие из них требования и методы их решения рассматриваются в последующих главах.

Во второй главе процесс анализа текста предлагается рассмотреть как операцию

$$P_h \quad T_h(S_1, S_2, \dots, S_{r(h)}) \rightarrow R_h(k_1, k_2, \dots, k_{r(h)}), \quad h = 1, \dots, L, \quad (1)$$

где T_h - формальное описание одной из грамматических конструкций входного текста в виде правила расстановки в тексте лексем из некоторого подмножества S_i , $i = 1, \dots, r(h)$, $r(h) < n$; R_h - r -арное отношение между концептами k_i , являющимися семантическими аналогами лексем S_i исследуемого фрагмента входного текста; P_h - правило трансформации конструкции $T(\cdot)$ входного текста в $R(\cdot)$; L - мощность множества правил входного текста.

Аналогично, организацию синтеза выходного текста, предлагается организовать как переход

$$P'_h \quad R'_h(k_1, k_2, \dots, k_{r(h)}) \rightarrow T'_h(S_1, S_2, \dots, S_{r(h)}), \quad h = 1, \dots, L', \quad (2)$$

где T'_h - правило расстановки в тексте лексем из некоторого подмножества S_i , $i = 1, \dots, r(h)$, $r(h) < n$, представляющее определенную грамматическую конструкцию выходного текста; R'_h - r -арное отношение между концептами k_i , являющимися семантическими аналогами лексем S_i из синтезируемого фрагмента выходного текста; P'_h - правило трансформации конструкции $T(\cdot)$ и ее перехода в $R(\cdot)$ для выходного текста; L - мощность множества правил выходного текста.

Предлагаемое конкретное машинное представление правил P_h и отношений $r(h)$ позволило получить схему машинного перевода, обеспечивающую практически полную независимость конструкций и грамматик выходного языка от входного. Использование модификаций в работе схем рис. 1,2 создало предпосылки для предложения архитектуры программных средств (ПС) СМП (рис. 1) и функциональной схемы работы СМП (рис. 2), обеспечивающей:

- принцип синтаксической управляемости программ ядра переводчика, если в качестве управляющей компоненты выступает непроцедурное описание входного и выходного языков;
- независимость от грамматик входных и выходных языков;
- возможность наращивания ядра переводчика динамически загружаемыми модулями, составленными на специальном, проблемно-ориентированном для лингвистов языке программирования. Язык этот в основном предназначен для задания правил анализа текстов с контекстно-зависимой грамматикой;
- максимальную технологическую независимость процесса модификации программных средств СМП при их развитии.

Необходимость активизации программ, являющихся составляющими векторов $\pi 1_i$, $\pi 2_i$, Π , Π' , возникает при анализе определенных грамматических конструкций и концептов. Запрос на их активизацию поступает на коммутатор из ядра. Последний обеспечивает загрузку и интерпретацию этих программ.

Векторы Π и Π' обычно сильно разрежены и в процессе перевода имеют значительные уровни частот активаций. По этой причине для обеспечения

необходимых скоростных характеристик их реализацию целесообразнее организовать на профессиональном языке системного программирования в виде динамически загружаемых модулей.

В отличие от П и П' (в особенности для английского языка), количество программ типа p_1 и p_2 значительно больше (примерно 25% словарных статей информационных баз СМП английского языка нуждается в необходимости таких программ). Частота же их использования определяется частотой использования данной словарной статьи информационной базы СМП. Поскольку частоты использования таких программ не высоки и они составляются лингвистами – непрофессионалами в области программирования, то для составления текстов этих программ используется проблемно-ориентированный процедурный язык. Операторы языка разрабатываются, исходя из удобства их использования, лингвистами, обеспечивающими организацию контекстного анализа текстов.

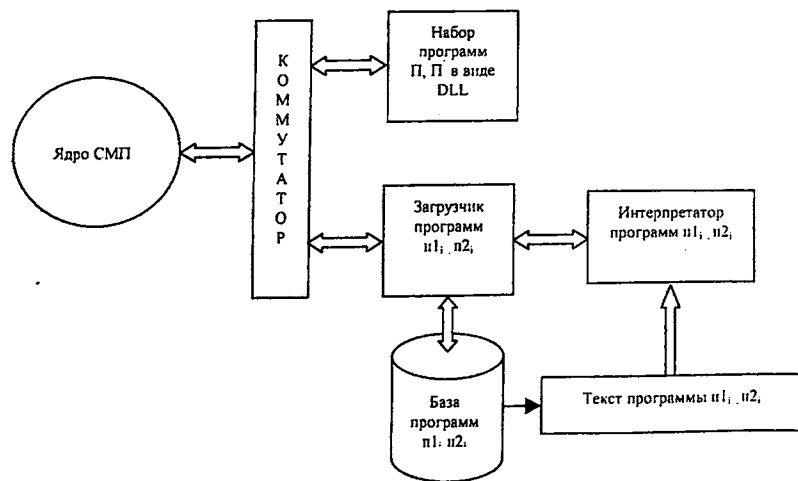


Рис. 1

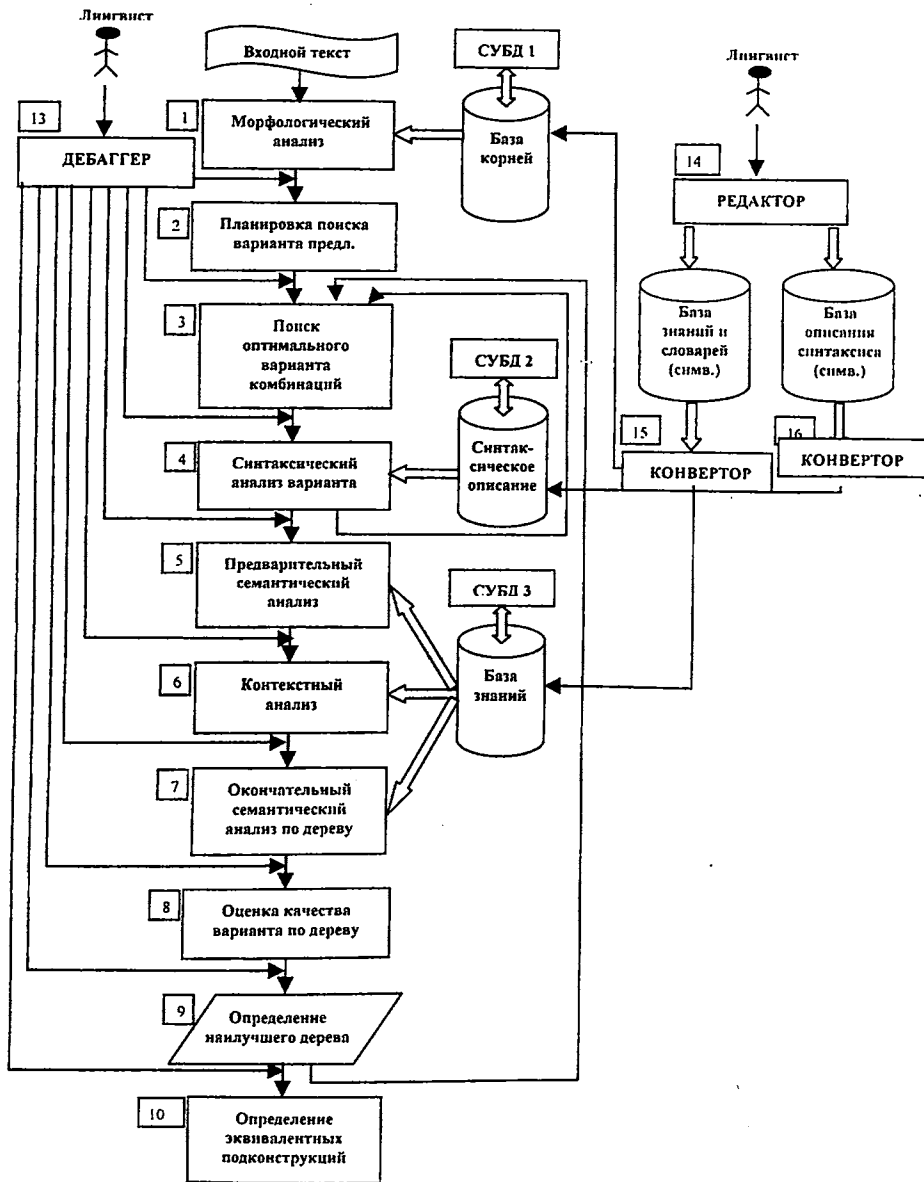
Вся необходимая информация для предлагаемой функциональной схемы СМП формируется лингвистами в виде двух баз, в первой из которых хранятся описания синтаксиса языков, а во второй - база знаний системы. Программная реализация схемы, приведенной на рис.2, позволяет создать ПС послышной архитектуры с максимально независимыми слоями. что наилучшим образом

способствует технологичности ее разработки. Функциональная независимость слоев позволяет при их проектировании и последующей разработке применить метод декомпозиции и исследовать их независимо. Данной проблеме посвящены последующие главы настоящей диссертационной работы.

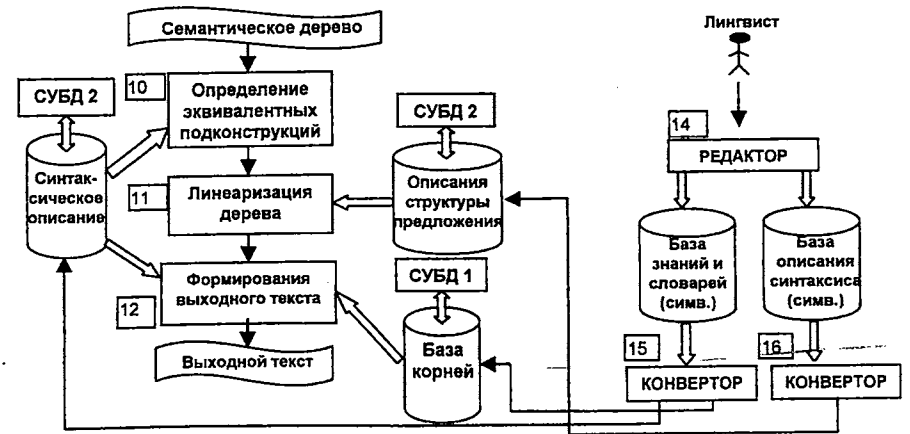
В третьей главе предлагаются методы организации информационных баз в системе машинного перевода. Реализация схемы рис. 1 сводится к поиску строк в T , фактически являющейся табличным представлением T_n . Найденные строки определяют возможный список синтаксически-семантических отношений r_{ij} по синтаксическим признакам трансформаций слов между соседствующими лексемами S_i и S_{i+1} .

Для рассматриваемых K_i и K_{i+1} предлагается определить соответствующие им списки I_i и I_{i+1} , куда входят обобщенные понятия в виде концептов предметной области. Принадлежность K_i к этим концептам требуется проверить в процессе анализа текста. Работы по выявлению списка r_{ij} предлагается выполнить в два этапа.

На этапе синтаксического анализа необходимо провести операцию фильтрации r_{ij} , вводя в набор ключей пару из соответствующих элементов списков I_i и I_{i+1} . Опыт разработки СМП показал, что выполнение на данном этапе частичного семантического контроля позволяет без существенного увеличения трудоемкости этапа синтаксического анализа резко уменьшить мощность списка r_{ij} . Тем самым уменьшаются затраты на реализацию конечного семантического анализа. Последний осуществляется на основе базы знаний, включающей концепты списка I . На этапе синтаксического анализа необходимо провести операцию фильтрации r_{ij} , вводя в набор ключей пару из соответствующих элементов списков I_i и I_{i+1} . Опыт разработки СМП показал, что выполнение на данном этапе частичного семантического контроля позволяет без существенного увеличения трудоемкости этапа синтаксического анализа резко уменьшить мощность списка r_{ij} . Тем самым уменьшаются затраты на реализацию конечного семантического анализа. Последний осуществляется на основе базы знаний, включающей концепты списка I .



а)



б)

Рис. 2. Алгоритм работы СМП

Требования, предъявляемые к базе знаний СМП, имеют свои особенности. К ним относятся: большое многообразие типов семантико-синтаксических отношений; частота выборки информации из БЗ (несравненно превышающая частоту модификации ее содержания) и, наконец, необходимость обеспечения прямого доступа к блокам данных, содержащим тела концептов K_i и K_{i+1} , по их идентификаторам. При таких требованиях использование традиционных реляционных баз становится неэффективным. Поэтому в работе предлагается использовать базу с сетевой логической и специальной физической организацией, обеспечивающей прямой доступ к блокам данных, представляющим концепты K_i и K_{i+1} . Поскольку с БЗ СМП общается большое количество лингвистов и ПС СМП, то предлагаются два типа физической организации для СМП: БЗ с символьным представлением для лингвиста и БЗ с бинарным представлением для ПС СМП. Перевод информации из первой базы во вторую осуществляется специальными конверторами (рис. 1).

При реализации предложенной структуры БЗ формирование семантического аналога входного предложения можно рассматривать как процесс выделения его из

сети базы знаний. т.е. сформированный семантический аналог является подграфом для графа БЗ.

Иллюстрацией сказанного служит приводимый пример предложения:

Ученик кушает свежеспеченный хлеб, купленный из магазина.

Семантический аналог и часть графа БЗ для данного предложения приведены на рис. 3 а, б.

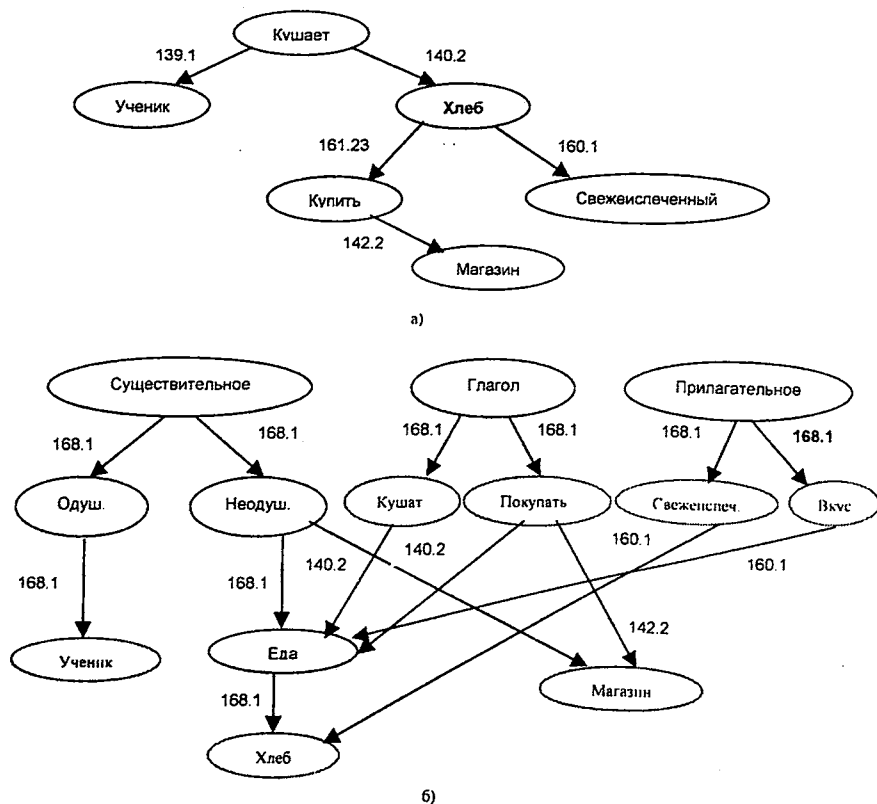


Рис. 3. Семантический аналог и часть графа БЗ для приведенного предложения

168.1 – код отношения “класс – подкласс”.

139.1 – код отношения “подлежащее-сказуемое”.

140.2 – код отношения “прямое дополнение”.

161.23 – код отношения, устанавливаемый между существительным и дополняющим его относительным прилагательным.

160.1 – код отношения, устанавливаемый между существительным и дополняющим его качественным прилагательным.

142.1 – код отношения между глаголом и его косвенным дополнением.

Задача выбора оптимального состава *K*; трудно формализуема, поэтому она решается инженерами по БЗ в основном на интуитивном уровне.

В средства управления БЗ предлагается встраивать программные блоки ведения статистических характеристик, позволяющие количественно оценить эффективность текущего варианта выбранной подструктуры БЗ и определить даты выдачи соответствующих рекомендаций инженерами по БЗ.

Для текстов с контекстно-зависимой грамматикой в состав описания языков необходимо внести программы анализа контекстов. Поскольку правила контекстной зависимости, а так же схемы контекстного анализа известны и задаются лингвистами, то метаязык этих правил должен быть контекстно-ориентированным. В связи с этим в работе предлагается процедурный метаязык. Основными объектами, с которыми манипулируют операторы этого языка, являются:

- переменные, представляющие позицию данного концепта в тексте программы;
- составные переменные, представляющие синтаксические параметры, характеризующие имена концептов, находящихся в заданных позициях;
- константы типа “метки”, “числа”, “слова”;
- переменные типа “множество”;
- списки концептов;
- деревья концептов.

В состав языка включены операторы:

- чтения и модификации параметров концептов;

- поиска в списке концептов концепта с заданными параметрами;
- удаления, включения или замены концептов в списке;
- проверки принадлежности конкретного концепта к определенному классу;
- навигации по ветвям графа;
- удаления, добавления и реструктуризации подграфов.

Наличие условных и безусловных операторов перехода позволяет строить разветвляющие программы, необходимые для организации контекстного анализа.

Скорость перевода во многом зависит от времени выборки информации из баз СМП. Последняя же определяется физической организацией этих баз. Учитывая особенности режимов эксплуатации СМП, в работе предлагается ввести две базы – базу корней слов и базу концептов, каждая из которых имеет свою специальную физическую организацию. Для БЗ используется сочетание прямых и инверсных форм, поддерживаемых в виде файлов ОС Windows. Для базы корней используется иерархическая система “Гора”. Получены оценки затрат объема памяти для организации базы корней, в зависимости от распределения символов в словах данного языка. Выработанные оценки позволят проектировщику в процессе разработки базы корней СМП определить эффективный вариант выбранной схемы “Гора”.

Для обеспечения нужного качества перевода в БЗ необходимо отобразить не менее 50 тысяч концептов. При таких мощностях БЗ имеет место резкое проявление омонимии, вследствие чего в процессе замены слова на концепты на одной и той же позиции слова появляется несколько претендентов - концептов. В процессе распознавания входного текста из указанных претендентов генерируются последовательности концептов, любая из которых является потенциальным смысловым аналогом предложения. Число ПК определяется величиной m^n , где m - среднее число претендентов – концептов на одно слово, n – среднее число слов предложения. Статистический анализ БЗ, содержащей порядка 50 тысяч концептов, показал, что значение m колеблется в среднем от трех до шести. При

средней длине предложения, содержащего от 8 до 12 слов, количество ПК достигает нескольких миллиардов.

Четвертая глава посвящена методам ускорения поиска смыслового аналога предложения.

Стратегия основана на определении оценок правдоподобия для каждого концепта – претендента в данном контексте. После морфологического анализа предлагается провести этап поверхностно - вероятностного анализа по следующей схеме:

1. К концепту приписывается список, элементы которого представляют время и область действия (ОД) данного концепта.
2. Композицией списков ОД распознанных предложений формируется список T_i , характеризующий тематику данных предложений. На основе же этого списка формируется T' , определяющий тематику текущего текста.
3. В качестве корреляционной меры принадлежности каждого из претендентов концепта к тематике анализируемого текста предлагается использовать выражение $k_0 = \alpha \cdot k + (1 - \alpha) \cdot k'$, где k - мера связанности данного концепта с тематикой, определяемой предшествующими предложениями. α - коэффициент, характеризующий возможную динамику обновления тематики при появлении новых предложений. Коэффициент k определяет меру связанности списка ОД текущего концепта со списком T .

Вычисляя претенденты k_0 и нормируя их, получаем искомые значения P_i - вероятности претендентов на данный концепт.

Конечная оценка правдоподобия получается после сочетания найденного значения P_i с оценкой синтаксически – семантической связанности данного слова с другими словами текущего предложения. На основе P_i формируется функция распределения претендентов для каждого слова, которая используется при построении плана поиска оптимальной ПК. Статистический анализ степени связанности двух произвольных i – го и j – го слов показал, что функция F вероятности наличия отношений между ними $P(i,j) = F(i-j)$ имеет явно выраженную

гиперболическую форму. При генерации ПК это обстоятельство позволяет обойти процесс полного перебора, реализуя лишь генерацию вариантов ПК - так называемых частично пересекаемых ограниченных отрезков перебора. Для обеспечения условия достижения минимума риска потерь при оптимальном решении предлагается алгоритм определения таких отрезков.

Испытание предложенной схемы планирования показало, что если предусмотренный ресурс выбрать пропорциональным общему количеству претендентов концептов на слова текста с коэффициентом пропорциональности, равным пяти, то составленный план поиска позволит сгенерировать оптимальный вариант ПК с вероятностью, превышавшей 97%.

Семантический анализ организуется путем поиска отношений между парами- концептами БЗ, представляющими слова предложения. Однако установить всевозможные отношения между концептами БЗ невозможно, поскольку тогда БЗ становится слишком громоздкой. В связи с этим при формировании БЗ используется лишь часть этих отношений, в то время как поиск не фиксированных в базе отношений осуществляется путем поиска путей (траекторий) между заданными концептами. В этих путях учитываются существующие в БЗ отношения, а найденные пути проходят через другие концепты БЗ. При поиске наикратчайших из этих путей в процессе навигаций по сети БЗ ставятся определенные ограничения на типы отношений и на факты возможных вариантов их соседства в сети.

В работе предлагается алгоритм определения наикратчайшего пути, основанный на подходе, при котором каждому i -му узлу сети БЗ ставится в соответствие некий узел в n - мерном пространстве. Тогда искомым наикратчайшим путем является наикратчайшим геометрическим путем. В работе предлагаются рекомендации по выбору размерности пространства, расстановки узлов в пространстве и алгоритма поиска пути в пространстве.

В пятой главе приводится архитектура разработанного программного обеспечения. Описан конкретный проблемно-ориентированный язык, предназначенный для организации контекстного анализа текста. Структура языка и

состав операторов ориентированы на использование лингвистами- непрофессионалами в области программирования. В процессе функционирования основного ядра переводчика предлагаемые механизмы обеспечивают динамическое подключение этих программ.

На конкретных примерах переводов, приведенных в главе в виде списка предложений, показано, что системой реализуется смысловой перевод основных типов конструкций переводимых предложений.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ РАБОТЫ

1. Предложено формальное описание процесса интеллектуального машинного перевода, предполагающее распознавание смысла входного предложения. Схема позволяет реализовать полный смысловой перевод, одновременно обеспечивая практически полную независимость конструкций и грамматик выходного языка от входного.
2. Основываясь на выбранной стратегии перевода, предложена послойная архитектура программных средств. Логическая и информационная организации физических баз обеспечивают:
 - максимальную независимость слоев, что позволяет организовать технологическую независимость процесса модификации программных средств СМП при их развитии;
 - принцип синтаксической управляемости программ ядра переводчика, когда в качестве управляющей компоненты выступает непроцедурное описание входного и выходного языков.
3. Предлагается схема организации поверхностно-вероятностного анализа, позволяющая определить достоверность претендентов - концептов на данное слово в зависимости от контекста.
4. Предлагается метод формирования плана генерации вариантов ПК частичного перебора с минимальным риском потери оптимального решения.

5. Предложены метод и алгоритм определения наикратчайших путей в семантической сети базы знаний для организации конечного семантического анализа.

Основные результаты диссертации представлены в следующих работах:

1. Манукян А., Хачатрян В., Манукян К., Аракелян Г. Организация этапа семантического анализа в системах машинного перевода // Сборник материалов годичной научной конференции ГИУА. -Ереван, 2001.-Т. 1. –С. 304-305.
2. Манукян Э., Манукян С., Манукян А., Саядян А. Основы построения интеллектуальной системы машинного перевода текстов // Сборник материалов годичной научной конференции ГИУА. -Ереван, 2001.-Т. 1. –С. 306.
3. Манукян Э., Манукян А. Определение эксплуатационных характеристик словарных баз систем машинного перевода // Сборник материалов годичной научной конференции ГИУА. -Ереван, 2002.-Т. 1. –С. 283-285.
4. Манукян Э., Манукян С., Манукян А., Саядян А. Об организации процесса самообучения в системе интеллектуального машинного перевода // Сборник материалов годичной научной конференции ГИУА. -Ереван, 2002. -Т. 1. –С. 282-283.
5. Manukyan E., Manukyan A., Manukyan S., Manukyan K. Constructive Approach to the Selection of Metalanguage for Control of the Linguistic Process of Machine Translation // In Proc. Of CSIT-2003 Conference. -Yerevan, Armenia. - P. 251-253.
6. Palyan A., Manukyan A. XML Documents Similarity Evaluation Method // In Proc. Of CSIT-2003 Conference. -Yerevan, Armenia. -P. 238-241.

7. Манукян А. Метод построения плана поиска “Графа грамматического разбора” предложения // Сборник материалов годичной научной конференции ГИУА. -Ереван, 2004.-Т. 1. –С. 387-390.
8. Манукян А., Назарян Г. Язык для организации контекстного анализа в системе машинного перевода // Сборник материалов Международной молодежной конференции информационных технологий. -Ереван, 2005. -С. 167-170.
9. Манукян Э., Манукян А. Об организации семантического анализа при машинном переводе // Известия НАН РА и ГИУА. Сер ТН. -2005. –Т 58. N3. -С. 579-584.

15.05.2014

ՄԱՆՈՒԿՅԱՆ ԱՐԱՄՍՈՒՐԵՆԻ

ԲԱՐՁՐ ՇԱՀԱԳՈՐԾՈՂԱԿԱՆ ԲՆՈՒԹԱԳՐԵՐ ԱՊԱՀՈՎՈՂ ՄԵՋԵՆԱՅԱԿԱՆ
ԹԱՐԳՄԱՆՈՒԹՅԱՆ ԿԱԶՄԱԿԵՐՊՄԱՆ ՄԵԹՈԴՆԵՐ

ԱՄՓՈՓԱԳԻՐ

Աշխատանքը նվիրված է ավտոմատացված թարգմանչական համակարգերի մշակման մեթոդներին, որոնք ապահովում են թարգմանության ներմեքենայական գործընթացների հարմարեցումը, թարգմանչի շահագործման արտաքին պայմանների փոփոխման դեպքում՝ բարձր շահագործողական բնութագրեր ստանալու նպատակով:

Աշխատանքի հիմնական արդյունքներն են.

1. Մշակված է թարգմանվող տեքստերի մոտավոր-հավանականային անալիզի կազմակերպման սխեմա, որը թույլ է տալիս որոշել թարգմանվող տեքստի մոտավոր թեման և ճշտել տվյալ բառին հավակնորդ կոնցեպտների հավաստիության աստիճանը:
2. Մշակված է թարգմանվող նախադասության բառերին հավակնող կոնցեպտների ցուցակներից այդ նախադասության իմաստին հավակնող կոնցեպտների շարանի ձևավորման սխեմա, որը թույլ է տալիս խուսափել բոլոր տարբերակների դիտարկման անհրաժեշտությունից և տրված ժամանակային սահմանափակումների համար ապահովել օպտիմալ շարվածքի կորստի ռիսկի նվազագույն արժեք:
3. Առաջարկված են գիտելիքների բազայի սեմանտիկ ցանցում ամենակարճ ճանապարհների որոնման մեթոդ և ալգորիթմ, որոնք օգտագործվում են թարգմանության գործընթացի վերջնական սեմանտիկ վերլուծության փուլում:
4. Մշակված է մետալեզու, որը կողմնորոշված է կոմպոստ-կախյալ քերականություն ունեցող տեքստերի վերլուծության համար:

