Институт проблем передачи информации РАН

ЛАБОРАТОРИЯ № 15

Лаборатория компьютерной лингвистики

Заведующий лабораторией: д.филол.н., Богуславский Игорь Михайлович
Тел.: (095) 299-49-27; Email: bogus@iitp.ru

Ведущие ученые лаборатории:

академик, д.филол.н.	Апресян Ю. Д.	к.ф.-м.н.	Цинман Л. Л.
д.филол.н.	Санников В. З.		Григорьев Н. В.
к.филол.н.	Григорьева С. А.		Крейдлин Л. Г.
к.филол.н.	Иомдин Л. Л.		Лазурский А. В.
к.ф.-м.н.	Митюшин Л. Г.		Фрид Н. Е.

НАПРАВЛЕНИЯ ИССЛЕДОВАНИЙ

Основной научной проблематикой лаборатории является функционирование естественного языка в качестве средства передачи информации. Фундаментальные исследования, проводимые в лаборатории, направлены на разработку полной действующей формальной модели языка типа "Смысл ⇔ Текст". Модель должна имитировать языковое поведение человека, т.е. его способность производить тексты на естественном языке и понимать их. Компьютерная версия модели, разрабатываемая в Лаборатории, имеет вид полифункционального многоязычного процессора, известного под условным наименованием ЭТАП. ЭТАП состоит из морфологических и комбинаторных словарей рабочих языков и наборов правил. В идеале правила во взаимодействии со словарями должны имитировать языковое поведение человека при производстве и понимании тестов. Объединенные в определенные модули, они приобретают и прикладную функцию, а именно, обеспечивают функционирование ряда построенных в Лаборатории систем переработки текстов, таких, как машинный перевод, порождение русских текстов по смысловому заданию на языке UNL, перифразирование предложений на данном естественном языке и т. п. Будучи инструментом решения ряда практических задач в области переработки текстов на естественных языках, такие системы, с другой стороны, служат экспериментальным полигоном для корректировки лингвистических описаний и получения принципиально новых лингвистических знаний. В 2002 году работа Лаборатории была направлена на расширение и совершенствование функциональных возможностей системы ЭТАП. Демонстрационная версия системы доступна по адресу http://proling.iitp.ru.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ

1) Продолжалась работа над пополнением и развитием комбинаторных словарей русского и английского языков.

– Словники этих словарей доведены до объема около 60000 единиц каждый, что соответствует объему крупных традиционных двуязычных словарей общего назначения.

– Лексикографическая информация в разных зонах словарей была пополнена и откорректирована. Наиболее существенной корректировке подверглись а) словарные статьи многозначных слов, б) модели управления, в) зона лексических функций. Основанием для пополнения и корректировки послужили, в частности, экспериментальные материалы, полученные в ходе функционирования разработанных в Лаборатории прикладных систем.

– Разработан модуль полуавтоматического ввода фразеологических единиц в комбинаторные словари.

2) Продолжалась работа над пополнением и развитием морфологических словарей русского и английского языков.

– Словари пополнялись за счет географических названий и собственных имен.

– Продолжалась работа по введению в морфологические словари композитов (типа квази- и quasi-) и компонентов сложных слов.

3) Начаты исследования по теме "Разработка теории и словаря глагольного управления для целей автоматического анализа и синтеза текстов на русском языке". Получены следующие результаты:

– Разработана новая лексико-семантическая теория глагольного управления, которая сводится к следующим тезисам: а) Глагольное управление отчасти семантически мотивировано, отчасти не мотивировано; б) мотивированные управляющие свойства глаголов определяются двумя семантическими факторами – системообразующими смыслами ‘действие’, ‘процесс’, ‘состояние’, ‘свойство’, ‘акциональность’, ‘цель’ и др., входящими в состав значений предикатных лексем, и самими их толкованиями; в) немотивированные управляющие свойства определяются индивидуальными лексическими особенностями глаголов.

– Составлен словник глаголов, подлежащих включению в словарь, объемом около 6000 единиц.

4) Разработан модуль перифразирования предложений русского языка на основе аппарата лексических функций; в ходе этой работы были получены следующие результаты:

– Разработан новый аппарат лексических функций, состоящий из новой системы определений лексических функций, зоны лексических функций в комбинаторных словарях и трех наборов правил: правил распознавания лексических функций в предложении, правил сведения предложений к их каноническому виду и правил собственно перифразирования. Этот аппарат был встроен в систему ЭТАП в качестве особого модуля.

– Осуществлена комплексная отладка системы перифразирования на компьютере и проведена большая серия экспериментов по перифразированию.

– Начат анализ экспериментального материала, позволивший не только отладить фрагменты лингвистических баз знаний, содержавших ошибки, но и серьезно продвинуться в понимании фундаментальных свойств языка.

5) Продолжалась работа по созданию русской подсистемы для программы многоязычной коммуникации на базе Универсального сетевого языка (UNL). Эта программа, разрабатываемая большим международным консорциумом под эгидой ООН, ставит своей целью обеспечить для пользователей Интернета возможность получать и распространять информацию на родном языке. Познакомиться с этой программой можно на сайте www.undl.org. Получены следующие результаты.

– Разработана первая версия модуля перевода с русского языка на UNL.

– Проведено совершенствование модуля перевода с языка UNL на русский. Этот модуль доступен по адресу www.unl.ru.

Институт проблем передачи информации РАН

– Проведен анализ спецификации языка UNL и разработаны предложения по ее совершенствованию.

6) Продолжалась работа по построению аннотированного корпуса русских текстов.

– Проведена аннотация 4000 предложений корпуса.

– Усовершенствовано программное обеспечение комплекса "Рабочее место аннотатора": разработана программа, которая либо сама исправляет обнаруженные в тексте ошибки, либо находит формальные дефекты морфо-синтаксических структур и выдает человеку список ошибок.

– Фрагмент корпуса передан компании Яндекс для совместной разработки поисковой системы и предоставления открытого телекоммуникационного доступа к корпусу через Интернет.

7) Разработан и интегрирован в лингвистический процессор ЭТАП оригинальный комбинированный алгоритм синтаксического анализа. В нем правила синтаксического анализа в процессе разрешения языковой неоднозначности динамически взаимодействуют со специально разработанным статистическим модулем, который приписывает веса гипотетическим синтаксическим связям на основе данных синтаксически размеченного корпуса текстов.

8) Построен первый образец принципиально новой интерактивной системы разрешения лексической неоднозначности для машинного перевода и других приложений. Система основана на динамическом взаимодействии алгоритма анализа с интуицией пользователя-человека, не имеющего никакой специальной лингвистической подготовки. Получив от пользователя ответы на предложенные ему простые вопросы (выбрать синоним данного слова, его перифразу, упрощенное толкование и т. п.), система обеспечивает стопроцентное различение лексической неоднозначности. С помощью этой системы обработано 20000 слов комбинаторного словаря, для которых выявлена структура полисемии и построены диагностические описания.

9) Разработаны лингвистические средства разрешения лексико-грамматической неоднозначности.

10) Разработана программа идентификации и маркировки в тексте именных групп, обозначающих лиц.

11) Обеспечена поддержка работы процессора ЭТАП с символами различных кодовых таблиц.

12) Разработана новая система HELP для процессора ЭТАП.

ГРАНТЫ:

- **Российский фонд фундаментальных исследований (№ 01-06-80453):** "Разработка комбинированного алгоритма синтаксического анализа для лингвистического процессора ЭТАП-3".
- **Российский фонд фундаментальных исследований (№ 01-07-90405):** "Создание аннотированного корпуса русских текстов (вторая очередь)".
- **Российский фонд фундаментальных исследований (№ 02-06-80085):** "Разработка интерактивной системы разрешения лексической неоднозначности для машинного перевода и других приложений".
- **Российский фонд фундаментальных исследований (№ 02-06-80106):** "Разработка теории и словаря глагольного управления для целей автоматического анализа и синтеза текстов на русском языке".

ПУБЛИКАЦИИ В 2002 г.

1. Апресян Ю.Д. Об одной закономерности устройства семантических систем // Проблемы семантического анализа лексики. Тезисы докладов международной конференции. М., 2002. С. 6-9.
2. Апресян Ю.Д. Взаимодействие лексики и грамматики: лексикографический аспект // Русский язык в научном освещении. 2002. № 3. С. 10-29.
3. Апресян Ю.Д. Новый объяснительный словарь синонимов русского языка: ход работы и результаты // Вестник Российского гуманитарного научного фонда. 2002. № 3. С. 87-99.
4. Апресян Ю.Д., Цинман Л.Л. Формальная модель перифразирования предложений для систем переработки текстов на естественных языках // Русский язык в научном освещении. 2003. № 4.
5. Богуславский И.М. «Сандхи» в синтаксисе: загадка *уже не* // ВЯ. 2002. № 5. С. 19-37.
6. Богуславский И.М., Григорьев Н.В., Григорьева С.А., Иомдин Л.Л. Разработка синтаксически размеченного корпуса русского языка // Доклады научной конференции "Корпусная лингвистика и лингвистические базы данных". С.-Пб.: изд-во С.Петербургского университета, 2002. С. 40-50.
7. Иомдин Л.Л. Уроки русско-английского (из опыта работы системы машинного перевода) // Труды Международного семинара Диалог'2002 по компьютерной лингвистике и ее приложениям. М.: Наука, 2002. Т. 2. С. 234-244.
8. Иомдин Л.Л., Сизов В.Г., Цинман Л.Л. Использование эмпирических весов при синтаксическом анализе // Обработка текста и когнитивные технологии. Казань: Отечество, 2002. № 6. С. 64-72.
9. Фрид Н.Е. Употребление настоящего исторического и прошедшего времен в спонтанной устной речи // Международная школа по лингвистической типологии и антропологии. Материалы лекций и семинаров. М.: РГГУ, 2002. С. 268-269.
10. Apresjan Ju.D. Principles of Systematic Lexicography // Lexicography and Natural Language Processing. A Festschrift in Honour of B.T.S. Atkins. Marie-Hélène Corréard (ed), Euralex 2002. P. 91-104.
11. Apresjan Ju.D., Boguslavsky I.M., Iomdin L.L., Tsinman L. L. Lexical Functions in NLP: Possible Uses. – In: Computational Linguistics for the New Millenium: Divergence or Synergy? Proceedings of the International Symposium held at the Ruprecht-Karls-Universität Heidelberg, 21-22 July 2000. Manfred Klenner / Henriëtte Visser (eds.) Frankfurt am Main, 2002. P. 55-72.
12. Boguslavsky I., Chardin I., Grigorjeva S., Iomdin L. et al. Development of a dependency treebank for Russian and its possible applications in NLP // Proceedings of the Third International Conference on Language Resources and Evaluation (LREC-2002), v. III, Las Palmas. P. 852-856.
13. Boguslavsky I. Some lexical issues of UNL // Proceedings of the First International Workshop on UNL, other interlinguas and their applications, Las Palmas, 2002. P. 19-22.
14. Iomdin L., Sizov V., Tsinman L. Utilisation des poids empiriques dans l'analyse syntaxique: une application en Traduction Automatique // META. 2002. V. 47. No. 3. P. 351-358.
15. Апресян Ю.Д. Остановка движения как симптом внутреннего состояния: синонимический ряд *замереть* // Сборник в честь В. Н. Сидорова (в печати).

16. Апресян Ю.Д. Системность лексики: семантические парадигмы и семантические альтернации // В сборник в честь С. Кароляка (в печати).
17. Апресян Ю.Д. Акциональность и стативность как сокровенные смыслы (охота на *оказывать*) // Сборник статей к 70-летию Н.Д. Арутюновой. М.: Языки русской культуры, 2003 (в печати).
18. Апресян Ю.Д. Трактовка вида в словаре: правила, тенденции, лексикализация // Сборник в честь проф. Лемана, Гамбург (в печати).
19. Апресян Ю.Д. Об одной закономерности устройства семантических систем // Пятые Шмелевские чтения (в печати).
20. Апресян Ю.Д. Принципы организации центра и периферии в лексике и грамматике // Сборник в честь В.С. Храковского (в печати).
21. Богуславский И.М. Замечания об актантной структуре адвербиальных дериватов (в печати).
22. Богуславский И.М. Часть – Целое – Признак: заметки о сфере действия кванторных слов (в печати).
23. Иомдин Л.Л. Идея и цель: об одном типе русских связочных предложений // Сборник статей к 70-летию Н.Д. Арутюновой. М.: Языки русской культуры, 2003 (в печати).
24. Apresjan Ju.D., Boguslavsky I.M., Iomdin L.L., Tsinman L.L. Lexical Functions in ETAP-3 // Сборник в честь 70-летия И.А. Мельчука (в печати).