ГЛАВА 4. ПРИКЛАДНЫЕ ВОЗМОЖНОСТИ СИНТАКСИЧЕСКИХ ПРОЦЕССОРОВ В СИСТЕМАХ МАШИННОГО ПЕРЕВОДА И АВТОМАТИЧЕСКОЙ ОБРАБОТКИ ТЕКСТОВ

В настоящей главе приводятся технические характеристики и оценка качества методик, разработанных для систем морфологического и предсинтаксического анализа и для процессоров синтаксической сегментации; дается краткое описание прикладных систем АОТ и МП, в которых были внедрены и опробованы предложенные методики; рассматриваются дальнейшие перспективы использования процессоров синтаксической сегментации в системах АОТ.

Качество методики морфологического анализа без словаря, разработанного в НТЦ «Система», оценивалось по следующим показателям :

- отношение количества ошибочно построенных гипотез к общему количеству полученных основ;
- минимальное количество словоформ одной лексемы, достаточное для гарантированного формирования правильной основы.

Общая скорость работы системы автоматической индексации БД характеризуется двумя показателями:

- скорость обработки текстов морфологическим анализом в процессе автоматического построения грамматического словаря основ;
- скорость индексации текстов с использованием построенного словаря основ.

Результаты испытаний методики морфологического анализа НТЦ «Система»:

Характеристика	Значение
Отношение количества ошибочно построенных гипотез к общему количеству полученных основ, [%]	5
минимальное количество словоформ одной лексемы, достаточное для гарантированного формирования правильной основы[1]	3
Скорость обработки в процессе построения словаря основ [Мб/ч]	6
Скорость индексации текстов с использованием построенного словаря основ [Мб/ч]	100

Полученное в результате тестирования отношение количества ошибочно построенных гипотез к общему количеству полученных основ показывает

[1] В 80% случаев достаточно 2 словоформ одной лексемы.

приемлемую погрешность метода. Заметим, что погрешность метода уменьшается при накоплении информации, т.е. увеличении выборки (количества словоформ для одной лексемы), которое зависит от объема «прочитанных» системой текстов. Скорость обработки существенно зависит от программной реализации методики. Имеющаяся программная реализация метода может быть значительно оптимизирована для дальнейшего промышленного использования. Морфологический анализатор без словаря был внедрен в первую версию ИПС законодательной БД, созданной на платформе Oracle 7.3 в НТЦ «Системы».

Качество морфологического анализа с использованием лексикона зависит от двух параметров:

- объем словаря – число включенных в словарь лексем;
- морфологическое покрытие – процент найденных слов в лексиконе в процессе анализа произвольного корпуса текстов.

Для морфологического компонента проекта Диалинг объем словаря составляет 165 тысяч лексем (в это число входят имена собственные и географические названия), покрытие – 98%. Скорость анализа достигает 200 Мб/ч.

Процессор LxPlatform 3.5 состоит из трех основных модулей: stemmer – морфанализ словоформы; tagger – снятие морфологической омонимии; np-grouper – выделение NP составляющих из текста. Приведем показатели оценки качества для модулей LxPlatform и скоростные характеристики:

Модуль	Характеристика	Значение
Stemmer	Объем словаря, [тыс. лексем]	150
Stemmer	Морфологическое покрытие, [%]	96
Stemmer	Скорость обработки текста, [Мб/ч]	450
Tagger	Отношение количества правильно выбранных для омонимичных словоформ лемм к общему количеству омонимичных словоформ, [%]	99
Tagger	Отношение количества правильно выбранных для омонимичных словоформ усеченных грамматических помет к общему количеству омонимичных словоформ, [%]	94,5
Tagger	Скорость обработки текста, [Мб/ч]	430
np-grouper	Отношение количества правильно построенных NP к общему количеству выделенных NP из текста, [%]	98,3
np-grouper	Скорость обработки текста, [Мб/ч]	400

Технология LxPlatform 3.5 была доведена до промышленного использования и успешно внедрена в системы АОТ отдела разработки Inxight.

Система АОТ Inxight	Функционал системы
Murax	Система, позволяющая расширить интеллектуальные возможности поисковой машины (АИПС), состоит из двух частей: (1) Concept Linker строит дерево концептов для множества индексируемых документов, где каждый концепт является частотной синтаксической составляющей NP (таким образом, выделение именных групп является центральной частью концептуального анализа документов в

	Murax); (2) Similarity Search позволяет находить для исходного текста родственные (похожие) документы.
Categorizer	Проводит автоматическую классификацию входящего информационного потока (документов) по определенной заранее таксономии. Классификатор необходимо обучать на таксономии заданной предметной области. Для полноценного обучения в среднем требуется 30 документов на каждую категорию.
Smart Discovery	Позволяет создавать в полуавтоматическом режиме таксономию предметной области на заданном массиве документов.

Каждое из вышеприведенных приложений использует LxPlatform в качестве ядра программы, подвергая результаты обработки текстов, полученных модулями tagger и np-grouper, дополнительному вероятностно-статистическому анализу.

Оценка качества работы синтаксического процессора определяется парой «точность (уровень ошибок в построенных синтаксических структурах предложений), полнота (степень покрытия текста синтаксическими связями, или связность графа предложения)».

Приведем показатели оценки качества и скорости для синтаксического анализа группы Диалинг[2]:

Характеристика	Значение
Отношение количества правильно сегментированных сложных предложений к общему количеству сложных предложений в тексте, [%]	78
Отношение количества правильно построенных синтаксических групп к общему количеству построенных групп в предложениях текста, [%]	97
Отношение количества правильно выбранных с использованием системы весов МИ сегмента к общему количеству выбранных МИ сегментов в предложениях текста, [%]	95
Покрытие: отношение количества слов, вошедших в состав синтаксических групп, к общему количеству слов в тексте, [%]	79
Скорость синтаксического анализа, [слов/сек.]	350

Синтаксический процессор является компонентой системы МП Диалинг. Результаты синтаксического анализа поступают на вход семантического модуля [А. Сокирко, 2001] системы, использующий в процессе работы русский общесемантический словарь (РОСС). Семантика берет от синтаксического компонента границы построенных сегментов, лучшие МИ сегментов, получившие максимальный вес, и только те синтаксические группы, которые имеют высокую точность построения и не требуют дополнительной семантико-синтаксической проверки через РОСС. Таким образом, центральной функцией синтаксического анализатора в составе системы МП Диалинг является сегментация сложного предложения и выбор МИ внутри сегментов (т.е. снятие грамматической омонимии).

[2] Все значения характеристик приводятся для состояния проекта Диалинг июня 2001г.

Синтаксический процессор группы Диалинг используется компанией «ВААЛ» (www.vaal.ru) для создания психологических методик анализа предвыборных, социологических и политических текстов [И. Ножов, 2002].

В Исследовательском центре искусственного интеллекта ИПС РАН в Переславле-Залесском синтаксический анализ группы Диалинг интегрирован как компонент лингвистического процессора в систему автоматического извлечения информации из текстов на русском языке [Д. Кормалев, Е. Куршев и др., 2002].

Приведем показатели оценки качества и скорости для программной реализации синтаксической сегментации группы ОИС РГГУ:

Характеристика	Значение
Отношение количества правильно сегментированных сложных предложений к общему количеству сложных предложений в тексте, [%]	85
Отношение количества правильно построенных синтагм к общему количеству построенных синтагм в предложениях текста, [%]	98
Скорость анализа, [слов/сек.]	450

Необходимо отметить, что данный анализатор не рассчитан на промышленное использование и, следовательно, имеет ряд ограничений в процессе анализа. Так, в сегментацию не включена обработка вводных и уточняющих оборотов и приложений, так как анализ подобных конструкций является второстепенной и решаемой задачей, решение которой не влияет на общий механизм алгоритмов сегментации. Имеющаяся программная реализация синтаксической сегментации может быть значительно оптимизирована для дальнейшего промышленного использования. Оптимизация программы и реализация метода активизации омонимов на языке С++ способны существенно увеличить скорость анализа (в 7-10 раз). Являясь внутриуниверситетским проектом, настоящая программная реализация синтаксической сегментации служит доказательством работоспособности лингвистических алгоритмов, разработанных Т.Ю. Кобзаревой, и предложенных методов монтажа и активизации омонимов.

Процессор синтаксической сегментации группы ОИС используется в учебном процессе Института Лингвистики РГГУ для проведения лабораторных и семинарских занятий.

Создание эффективной системы сегментации сложного предложения позволит осуществлять поиск заданного образца в пределах одного сегмента, что существенно повысит качество работы интеллектуальных ИПС. Сегментационный анализатор, способный выделить сегмент простого предложения в составе сложного, может стать центральным звеном в программах автоматического реферирования текстов. Сегментация предложения - компонента полного

синтаксического анализа, без которой представляется невозможным полноценное решение задач извлечения информации из текстов, автоматической кластеризации информационного потока и машинного перевода.

ЗАКЛЮЧЕНИЕ

Сформулируем основные результаты исследования:

- проведено сравнение существующих подходов к проектированию синтаксического анализа языка: опирающейся на лексикализм унифицирующей грамматики (HPSG), конекстно-свободной грамматики (LinkParser) и модели поверхностного текстового процессора (STP);
- разработаны два метода автоматического синтаксического анализа предложения: метод активизации омонимов и рекурсивный метод монтажа сегментов. Метод активизации омонимов реализует в системе синтаксической сегментации принцип ленивых вычислений и позволяет избежать полного декартова произведения морфологических омонимов и повторного построения общих для омонимичных графов синтагм и сегментов, что существенно снижает число рассматриваемых структурных интерпретаций предложения. Метод монтажа, опирающийся на свойство сегментной проективности в естественном языке, является базовым механизмом для выделения и классификации сегментов в составе сложного предложения;
- метод монтажа и метод активизации омонимов лингвистически адекватны и универсальны (независимы от анализируемого ЕЯ). Адекватность понимается как соответствие модели процессора трем сформулированным принципам: описательному, объяснительному и эмулирующему;
- автоматическая синтаксическая система ОИС представляет собой не законченный промышленный продукт, а экспериментальную систему для отработки лингвистических решений;
- реализация системы ОИС позволила создать автоматическую синтаксическую сегментацию русского предложения без искусственных ограничений на анализ;
- разработан оригинальный метод прикладного морфологического анализа без использования словаря, опирающийся на построение леса деревьев морфологических гипотез и сравнение (корреляция) через матрицы

инцидентности полученных деревьев для дальнейшей унификации грамматических гипотез;

– экспериментально доказана возможность применение скрытых моделей Маркова для задачи снятия морфологической омонимии в русскоязычном тексте.

ЛИТЕРАТУРА

[D.Grune, C.Jacobs, 1990] D.Grune, C.Jacobs. Parsing Techniques. A practical guide. – Vrije Universiteit, Amsterdam, 1990.

[Г. Буч, 2000] Г. Буч. Объектно-ориентированный анализ и проектирование. – М.: «Издательство Бином», 2000.

[I. Sag, T. Wasow, 1999] Ivan A. Sag, Thomas Wasow. Syntactic Theory: A Formal Introduction. – Stanford University, 1999

[С. Бурлак, С. Старостин, 2001] С. А. Бурлак, С. А. Старостин. Введение в лингвистическую компаративистику. – Эдиториал УРСС, М., 2001.

[M. Boden, 1990] M. Boden. Artificial intelligence and images of man. // Perspectives From Artificial Intelligence, 1990.

[Xerox, 1999] Examples of Networks and Regular Expressions. // www.xrce.xerox.com/research

[И. Мельчук, 1999] И. А. Мельчук. Опыт теории лингвистических моделей «Смысл ⇔ Текст». - М., 1999.

[Ф. де Соссюр, 1999] Ф. де Соссюр. Курс общей лингвистики. – М., 1999.

[S. Oepen, K. Netter, 1997] S. Oepen, K. Netter. Test Suites for Natural Language Processing. // Linguistic Databases, CSLI Lecture Notes #77.

[D. Sleator, D. Temperley, 1991] D. Sleator, D. Temperley. Parsing English with a Link Grammar. – CMU-CS-91-196, School of Computer Science, Carnegie Mellon University, Pittsburg, 1991.

[D. Grinberg, J. Lafferty, 1995] D. Grinberg, J. Lafferty. A robust parsing algorithm for Link Grammars. - CMU-CS-95-125, School of Computer Science, Carnegie Mellon University, Pittsburg, 1995.

[XRCE MLTT, 1995] Application of Finite-State Networks. // www.xrce.xerox.com/research

[Н. Леонтьева, 1995] Н. Н. Леонтьева. «Политекст»: информационный анализ политических текстов. // НТИ, Сер.2, 1995, №4.

[Н. Сущанская, 1989] Н. Ф. Сущанская. Программный препроцессор для естественноязыковых интерфейсов. - Автореф. дисс. к.т.н. – К.: РИО ИК, 1989.

[С. Гладунов, О. Федяев, 2002] С. А. Гладунов, О. И. Федяев. Распознавание речи на основе нейросетевой аппроксимации фонем. // КИИ-2002. Труды конференции, т.1 – М., Физматлит, 2002.

[Я. Тестелец, 2001] Я. Г. Тестелец. Введение в общий синтаксис. – М., РГГУ, 2001.

[С. Эйзенштейн, 2000] С. М. Эйзенштейн. Монтаж. – М., 2000.

[Х. Дрейфус, С. Дрейфус, 1998] Дрейфус Х., Дрейфус С. Создание сознания vs моделирование мозга. //Аналитическая философия: Становление и развитие. М., 1998.

[Д. Серл, 1998] Серл Д. Мозг, сознание и программы. //Аналитическая философия: Становление и развитие. М., 1998.

[Х. Патнэм, 1999] Патнэм Х. Философия сознания. //М., 1999.

[ВИНИТИ, 1990] Итоги науки и техники: физические и математические модели нейронных сетей, том 1, М., изд. ВИНИТИ, 1990.

[А. Кибрик, 2001] Кибрик А.Е. Очерки по общим и прикладным вопросам языкознания. – УРСС, М., 2001.

[Э. Сепир, 1993] Э. Сепир. Избранные труды по языкознанию и культорологии. //М., 1993.

[В. Ингве, 1965] В. Ингве. Гипотеза глубины. //Новое в лингвистике. Вып. 4, М., 1965 – с. 126.

[Б. Страуструп, 1999] Б. Страуструп. Язык программирования С++. – М., 1999.

[Н.А.Еськова, И.Г.Бидер и др.] Н.А.Еськова, И.Г.Бидер и др. Формальная модель русской морфологии.

[С.О.Шереметьева, С.Ниренбург, 1996] Эмпирическое моделирование в вычислительной морфологии. //НТИ, №7, 1996.

[Г.Г.Белоногов, 1984] Г.Г.Белоногов. Итоги науки и техники. Серия “Информатика”, т.№8,1984г.

[J. Goldsmith, 1999] J. Goldsmith. Unsupervised Learning of the Morphology of a Natural Language. //University of Chicago, 1998.

[И. Ножов, 2000] Ножов И.М. Прикладной морфологический анализ без словаря. // КИИ-2000. Труды конференции – М.: Физматлит, 2000. Т.1. С. 424-429

[И. Ножов, 2000] Ножов И.М. Процессор автоматизированного морфологического анализа без словаря. Деревья и корреляция. //Диалог’2000. Труды конференции - Протвино, 2000. Т.2. С. 284-290.

[А.Зализняк, 1980] Зализняк А.А. Грамматический словарь русского языка - М.: Русский язык, 1980 г.

[Ф.Харари, 1973] Ф.Харари. Теория графов. - М., 1973.

[Lauri Karttunen, 1993] Lauri Karttunen. Finite-State Lexicon Compiler. //Technical Report. ISTL-NLTT, Xerox Palo Alto Research Center, Palo Alto, California, 1993.

[Lauri Karttunen, 1994] Lauri Karttunen. Constructing Lexical Transducers. //15th International Conference on Computational Linguistics. Coling 94, I, pages 406-411. August 5-9, 1994. Kyoto, Japan.

[Finite-State Network, 1995] Finite-State Network. // Xerox Research Center, Grenoble, www.xrce.xerox.com/research

[Ж.Г.Аношкина, 1995] Ж.Г.Аношкина. Морфологический процессор русского языка. //Альманах «Говор», Сыктывкар, 1995, с.17-23.

[Ч. Хоккетт, 1961] Ч. Хоккетт. Грамматика для слушающего. // Новое в лингвистике. Вып. 4, М., 1965 – с. 139.

[S. Oepen, J. Caroll, 2000] S. Oepen, J. Caroll. Parser engineering and performance profiling. // Journal of Natural Language Engineering # 6 (1), 2000.

[Т. Кормен и др., 2001] Т. Кормен, Ч. Лейзерсон, Р. Ривест. Алгоритмы, построение и анализ. – М., МЦНМО, 2001.

[G. Neumann, J. Piskorski, 2001] G. Neumann, J. Piskorski. A Shallow Text Processing Core Engine. – DFKI, Saarbruecken, 2001.

[Дж. Фридл, 2001] Дж. Фридл. Регулярные выражения. – СПб., 2001.

[А. Сокирко, 2001] А. В. Сокирко. Семантические словари в автоматической обработке текста (по материалам системы Диалинг). - Автореф. дисс. к.т.н. – М., 2001.

[Д. Панкратов и др., 2000] Д. В. Панкратов, Л. М. Гершензон, И. М. Ножов. Описание фрагментации и синтаксического анализа в системе Диалинг. // Техническая документация, www.aot.ru, 2000.

[Т.Ю.Кобзарева, 2002] Т.Ю. Кобзарева. Некоторые аспекты анализа сочинения при сегментации русского предложения (неоднозначности при появлении «матрешек»). // КИИ-2002. Труды конференции – М.: Физматлит, 2002. Т.1. С.192-198.

[И. Ножов, 2002] И.М. Ножов. Проектирование сегментационного анализатора русского предложения. // КИИ-2002. Труды конференции – М.: Физматлит, 2002. Т.1. С. 212-222.

[А. Белоусов, С. Ткачев, 2001] А. И. Белоусов, С. Б. Ткачев. Дискретная математика. - т.19, М.,2001.

[Т. Кобзарева и др., 2000] Т.Ю. Кобзарева, Д.Г. Лахути, И.М. Ножов. Сегментация русского предложения. // КИИ-2000. Труды конференции – М.: Физматлит, 2000. Т.1. С. 339-344.

[Т. Кобзарева и др., 2001] Т.Ю. Кобзарева, Д.Г. Лахути, И.М. Ножов. Модель сегментации русского предложения. // Диалог'2001. Труды конференции – Аксаково, 2001. Т.2. С. 185-194.

[Н. Вирт, 2001] Н. Вирт. Алгоритмы и структуры данных. – СПб., 2001.

[И. Ножов, 2002] Ножов И.М. Синтаксический анализ. //Компьютерра, № 21 (446), 2002.

[Д. Кормалев, Е. Куршев и др., 2002] Кормалев Д.А., Куршев Е.П., Сулейманова Е.А., Трофимов И.В. Извлечение данных из текста. Анализ ситуаций ньюсмейкинга. // КИИ-2002. Труды конференции, т.1 – М., Физматлит, 2002.

[И. Мельчук, 1997] Курс общей морфологии - Т.№1, М., 1997.