# Исследование методов извлечения информации из текстов с использованием автоматического обучения и реализация исследовательского прототипа системы извлечения информации

***Куршев Е. П., Кормалев Д. А., Сулейманова Е. А., Трофимов И. В.***

epk@epk.botik.ru, dk@conrad.botik.ru, yes@helen.botik.ru, igor@warlock-98.botik.ru

Переславль-Залесский, ИПС РАН

Задача извлечения информации из текста [8] заключается в автоматической обработке набора документов с целью выделения релевантных данных и представления их в структурированной форме. По глубине анализа текста и степени перехода от текста к модели предметной области технология извлечения информации (ТИИ) занимает промежуточное место между информационным поиском и гипотетической технологией понимания текстов. Извлечение информации может осуществляться в «слабом» или «сильном» смысле. Круг задач, решаемых системами первого типа, относится к так называемым «малым» задачам семантического анализа текстов [5], так как для их решения достаточен локальный контекст и ограниченный, локальный, синтаксический анализ. Результаты извлечения информации в «слабом» смысле и характер их представления несколько ограничивают возможности дальнейшего использования добытых из текста данных. Извлечением информации в «сильном» смысле мы назвали бы переход от базы текстовых фактов к такому их представлению, которое можно было бы использовать как интеллектуальный информационный ресурс, своего рода базу текстовых знаний.

Наши исследования были направлены на усовершенствование методов и расширение возможностей ТИИ, что позволило бы подойти вплотную к решению задачи извлечения информации в «сильном» смысле.

## Средства технологии извлечения информации

Для выражения знаний о предметной области в задачах извлечения информации используется два основных вида средств: правила, описывающие текстовые ситуации (контексты), и ресурсы знаний. ТИИ обычно использует модель текста, основанную на аннотациях [9], отличающуюся простотой и высокой степенью универсальности. Описание текстовых ситуаций выполняется при помощи правил на специальном языке. При использовании модели аннотаций используются языки, основанные на языке CPSL [7]. Правило CPSL состоит из двух частей: образец и действия, выполняемые при успешном сопоставлении образцу. Образец для сопо-

ставления — обобщение регулярного выражения, где в роли символов выступают аннотации.

**Развитие языка правил.** Нами был разработан расширенный диалект языка CPSL. Предлагаемые нами расширения [3] преследуют две цели: 1) обеспечить возможность описания более сложных контекстов и 2) снизить объем рутинной работы при создании системы правил за счет более компактного описания контекста. В число предложенных расширений языка правил входят развитая система типов данных для значений атрибутов аннотаций, логические метасимволы, метасимволы перехода, списки значений («микрословари»), опережающая и ретроспективная проверка, дополнительные квантификаторы [3].

**Усовершенствование ресурсов знаний.** Интерес представляют не столько системы собственно извлечения информации, сколько системы, обеспечивающие возможности аналитической обработки накопленной информации. Для достижения такого результата недостаточно лишь таксономии понятий предметной области, обогащенной атрибутивными или меронимическими связями (т. е. тезауруса), — требуется набор ресурсов знаний, всесторонне описывающих предметную область. В ходе изучения подходов к построению ресурсов знаний сформировалась многомерная классификация ресурсов по их характеру и назначению: «предметные–лингвистические», «априорные–полученные из текстов», «описание концептов–описание экземпляров» [6].

### Применение машинного обучения

Трудоемкость построения контекстных правил вручную высока, поэтому для облегчения разработки и настройки приложений желательно использовать средства автоматизированного создания правил. Желательно также было бы в автоматическом или автоматизированном режиме пополнять ресурсы знаний.

**Автоматическое построение правил.** По своей сути задача построения набора правил близка к задаче наполнения высокоточных специализированных словарей моделей управления. Получаемые правила должны обеспечить извлечение информации в «слабом» смысле: распознавание текстовой ситуации, выделение целевых объектов и их свойств, а также, возможно, идентификацию некоторых отношений. Мы выбрали индуктивный подход к построению правил. Обучение идет на основе размеченных примеров, при этом используется две основных операции: обобщение и специализация [2]. Для повышения точности правил и обеспечения достаточной степени обобщения используется двухфазный сценарий обучения [4]. Предварительные эксперименты показали перспективность такого подхода.

**Автоматизированное пополнение ресурсов знаний.** Вопрос интеграции произвольных извлеченных текстовых фактов в ресурсы знаний пока открыт. Более насущные задачи ставит перед нами необходимость оперативного «запоминания» и использования извлеченной из некоторого текста информации для анализа этого же текста. Минимальная интеллектуальность системы предполагает, что текстовое выражение однажды распознанного объекта в дальнейшем будет рапознаваться и без «диагностического» контекста. По мере анализа текста идет пополнение динамического словаря текста [1], в дальнейшем эта информация подтверждается или опровергается экспертом. Для полноценного использования динамического словаря нужно решить ряд проблем: обработка неизвестных морфологическому анализатору склоняемых слов, учет синтаксической модели, описывающей словоизменение распознанной конструкции.

Работа выполнена при поддержке РФФИ, проект № 05-01-00442а.

## Литература

[1] *Александровский Д. А., Кормалев Д. А., Кормалева М. С., Куршев Е. П., Сулейманова Е. А., Трофимов И. В.* Развитие средств аналитической обработки текста в системе ИСИДА-Т // КИИ-2006. Труды конференции. — Т. 2. — М.: Физматлит, 2006. — С. 555–563.

[2] *Кормалев Д. А.* Автоматическое построение правил извлечения информации из текста // 1-я межд. конф. «Системный анализ и информационные технологии» САИТ–2005. — Т. 1. — М.: КомКнига, 2005. — С. 205–209.

[3] *Кормалев Д. А., Куршев Е. П.* Развитие языка правил извлечения информации в системе ИСИДА-Т // Межд. конф. «Программные системы: теория и приложения», ИПС РАН, Переславль-Залесский, октябрь 2006 г. — Т. 1. — М.: Физматлит, 2006. — С. 365–377.

[4] *Кормалев Д. А.* Обобщение и специализация при построении правил извлечения информации // Конф. КИИ-2006. — Т. 2.— М.: Физматлит, 2006. — С. 572–579.

[5] *Леонтьева Н. Н., Семенова С. Ю.* Инструменты построения фрейма «ПЕРСОНА» // НТИ, Сер. 2. Информ. процессы и системы. — 2001. — № 8.

[6] *Сулейманова Е. А.* Классификация ресурсов знаний в системе извлечения информации из текста // ММРО-13 (наст. сб.). — 2007. — С. **??**–**??**.

[7] *Appelt D. E.* The Common Pattern Specification Language: Technical report. — SRI International, Artificial Intelligence Center, 1996.

[8] *Appelt D. E., Israel D. J.* Introduction to Information Extraction. Tutorial // 16th Int'l. Joint Conf. on Artificial Intelligence IJCAI'99, Sweden, 1999.

[9] *Grishman R.* TIPSTER Text Architecture Design. Version 3.1. — New York: NYU, 1998.