КОМПЬЮТЕРНАЯ МАТЕМАТИКА

УДК 680.3.06

О.М. ДЕМСКАЯ-КУЛЬЧИЦКАЯ, В.Р. СЕМЕРЕНКО, Р.А. ЮЩЕНКО

МЕТОДЫ АВТОМАТИЧЕСКОЙ РАЗМЕТКИ ТЕКСТОВ НАЦИОНАЛЬНОГО КОРПУСА ЯЗЫКА

Методы автоматической разметки рассмотрены в контексте исследования текстов естественного языка, построения и поддержки Национального корпуса украинского языка. Представлен обзор основных подходов. Приведена схема распараллеливания обучающего метода автоматической разметки на кластере.

Введение. Актуальное направление компьютерной лингвистики - языки спецификаций, позволяющие структурировать и систематизировать взаимосвязи и формы представления информации посредством разметки текста. Их используют для спецификации внешнего представления информации (HTML) и для унификации структуры семантической составляющей (SGML, XML). Объединяя форму представления информации с ее содержанием, языки спецификаций удобны в исследовании текстов естественного языка [1-3].

Корпус - это набор размеченных текстов в электронной форме, представляющих язык на определенном этапе (или этапах) его существования и отображающих все многообразие жанров, стилей, территориальных и социальных вариантов языка. Трудоемкость разметки отобранных текстов очевидна, поскольку корпуса естественных языков отличаются структурой, обусловленной целью их создания, и могут содержать от десятков тысяч до десятков миллионов слов. Ряд математических методов позволяет частично или полностью формализовать львиную долю работ по разметке текстов. Основные подходы - обучающий (rule-based), статистический и трансформационный [1-6].

Любой корпус обеспечивает научные исследования лексической и грамматической структуры языка, а также тонких непрерывных процессов языковых изменений. Национальный корпус включает некоторое число

текстов конкретного языка, охватывающих максимально полный набор стилей, жанров и форм (устная, письменная), для изучения языка в динамике, определения происхождения и употребления слов и грамматических форм, изучения языковых качеств текстов. Источниками текстов, входящих в корпуса, для опубликованных книжных, журнальных и газетных текстов, как правило, являются выверенные электронные версии, предоставляемые издателями этих текстов.

По цели создания выделяют несколько типов корпусов. Корпус может включать один (monolingual), два (bilingual) и более (multilingual) языков. К национальным корпусам относят британский (British National Corpus), американский (American National Corpus), польский (Polski korpus narodowy), чешский (Czech National Corpus), русский (Национальный корпус русского языка). Отдельные корпуса изучают детскую речь (корпус CHILDES), речь неносителей языка (International Corpus of Learned English), устную речь, исторические тексты разных временных периодов. Существуют многоязычные корпуса текстов и параллельные или сравнительные корпуса, используемые в теории и практике перевода и составленные из нескольких аналогичных по структуре одноязычных корпусов текстов.

Структура текстов корпуса обусловлена характером решаемых задач, например, анализ отдельных слов, словосочетаний, групп слов (синтаксических или семантических), предложений, абзацев, и т. д. Корпус реализует:

- средство автоматизированной разметки;
- обработчик запросов к корпусу;
- корректировщик текстов для обеспечения адекватности возникающим требованиям.

Запрос - это строка, определяющая задачу поиска в корпусе. Поисковая машина, проанализировав запрос, производит поиск и формирует результат как набор отрывков из текстов, отвечающих требованиям запроса. Эти тексты сопровождаются информацией, такой как: название, автор, время написания и т.д.

С течением времени происходят морфологические, синтаксические, семантические изменения языка (включающие: появление новых слов, выход из употребления слов, появление новых значений существующих слов). Чтобы корпус продолжал быть адекватным предъявляемым требованиям необходимо постоянно корректировать его. Развитие корпуса происходит в нескольких направлениих, например, количественное пополнение корпуса текстами разных стилей, ранее в нем не представленных.

Эффективность корпуса определяется способностью к быстрой обработке запросов, точности и полноте результатов, поддержке одновременной работы нескольких пользователей. Эффективность достигается продуманным пользовательским интерфейсом, используемыми методами автоматической разметки текста и информационного поиска.

Построение корпусов славянских языков учитывает опыт корпусов романо-германской группы, однако славянские языки имеют свои особенности, они имеют завершенные и незавершенные формы глаголов, наличие множества

форм слов (например, насчитывается более 40 таких форм глагола «выйти», как «выйди», «вышедший», «вышел», «вышедшее» и т. п.).

Анализ текста на естественном языке зачастую предполагает многоуровневую разметку текста. Так, размеченный текст, полученный на одном уровне анализа, используется как исходный для разметки текста на более глубоком уровне. Синтаксически аннотированный корпус текстов строят в следующей последовательности: морфологический анализ, морфосинтаксическое тегиров^ние, поверхностный синтаксический анализ, глубокий синтаксический анализ, синтаксический анализ на уровне составных частей, синтаксический анализ на зависимости, синтаксически аннотированный текст. На ранних этапах структурирования текста посредством разметки решаемые задачи полностью формализуемы (например, морфологический разбор), углубленный анализ требует использования интеллектуальных средств и привлечения человека к решаемой задачи. К задачам интеллектуальной обработки относится соотнесение сегментов текста к заранее определенным систематизированным категориям, которые образуют схему. Такие категории могут иметь иерархическую структуру.

Обучающие методы активно используют в компьютерной лингвистике для двухэтапного анализа естественно-языковых текстов. На первом этапе, согласно словарю, с каждым словом анализируемого текста сопоставляется список возможных разметок. На втором этапе по правилам, сформированным в процессе обучения, из возможных вариантов разметки выбирается единственно правильная. Преимущество такого подхода - точность, недостаток - необходимость в большом числе правил, сформулированных вручную.

Обучающие методы автоматической разметки предполагают решение проблемы неопределенности наборов слов. В данной области ведутся активные исследования, среди которых: мультипликативный алгоритм коррекции весов, латентный семантический анализ, алгоритм трансформационного обучения, дифференциальные грамматики, метод списков решений, и множество байесовских алгоритмов классификации [3-7]. Здесь проблема устранения неопределенности наборов слов сформулирована так. В наборе слов все вхождения членов неопределенного набора заменяются маркером; каждый раз, когда система встречает маркер, она должна решить, какой из членов набора следует использовать в данном контексте. Обучающая система ориентирована на небольшое число вариантов неопределенности и на возможность определения этого варианта исходя из контекста. Контекст можно определять окном вокруг слова (оконно-ориентированные обучающие алгоритмы); непосредственными соседями слова (в обучающих системах с памятью), персептроном, байесовскими методами. Проблема устранения неопределенностей наборов слов охватывает проблемы: определение лексического значения слова (word sense disambiguation), распознавание части речи (part of speech tagging).

Статистический подход предполагает среди всех возможных разметок для слов предложения проставить самую вероятную последовательность.

Пусть дана последовательность слов $W = w_1, w_2, ..., w_n$. Рассчитаем наиболее вероятную последовательность тегов $T = t_1, t_2, ..., t_n$, максимизирующую $\arg\max(t_1, t_2, ..., t_n \mid w_1, w_2, ..., w_n)$. Применив закон Байеса, получаем:

$$P(t_1, t_2, ..., t_n \mid w_1, w_2, ..., w_n) = \frac{P(w_1, w_2, ..., w_n \mid t_1, t_2, ..., t_n) \cdot P(t_1, t_2, ..., t_n)}{P(w_1, w_2, ..., w_n)}.$$

Таким образом, необходимо найти $t_1, t_2, \ldots, t_n$, максимизирующие $P(w_1, w_2, ..., w_n \mid t_1, t_2, ..., t_n)$. Для этого упростим:

$$P(w_1, w_2, ..., w_n \mid t_1, t_2, ..., t_n) \approx \prod_{i=1}^{n} P(w_i, t_i) \text{ до } P(t_1, t_2, ..., t_n) \approx \prod_{i=1}^{n} P(t_i \mid t_{i-1}).$$

Задача свелась к нахождению $t_1, t_2, ..., t_n$, максимизирующих $\prod_{i=1}^{n} P(t_i \mid t_{i-1}) \times$ $\times P(w_i, t_i)$. Таким образом, для конкретного слова t_i необходимо найти $t_i = \arg\max_j P(t_j, t_{j-1}) P(w_i, t_j)$.

Значит, для нахождения наиболее вероятной разметки текста необходимо знать априорные вероятности разметки каждого слова $P(w_i, t_j)$ (такие вероятности известны в силу однозначности разметки большинства слов) и вероятности $P(t_j, t_{j-1})$, образующие цепь Маркова. Отсюда схему автоматической разметки байесовским методом можно изобразить рекуррентным соотношением:

$$P_1(t_j) = P(w_1, t_j),\ i = \overline{1, M},\ P_i(t_j) = \max_k (P_{i-1}(t_k) \cdot P(t_k, t_j) \cdot P(w_i, t_j)).$$

Основной недостаток байесовского подхода - сложность определения априорной вероятности соотнесения разметки каждому слову.

Автоматическая разметка украинских текстов на кластере. Для построения Национального корпуса украинского языка созданы словари морфологического разбора слов, включающие более миллиона словоформ (см. рис. 1, где пунктиром отмечено N узлов кластера). Корпус проходит этап формирования набора размеченных текстов. Ввиду объемов словарей и требования к скорости обработки текстов, задачу автоматической разметки целесообразно решать на кластере.

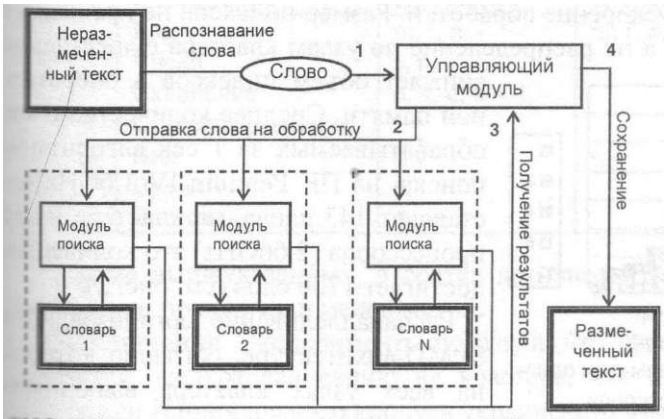

РИС. 1. Схема параллельной автоматической разметки

Вход втоматической разметки - неразмеченный текст на украинском языке, (см. рис. 1). Цифры возле управляющего модуля показывают последовательность обработки данных.

Обработка текста включает:

1) разбор текста и выделение слов;
2) передача слова подчиненным узлам кластера для обработки;
3) получение и объединение результатов подчиненных узлов;
4) представление результатов разбора в виде разметки и сохранение ее.

Для поиска и выделения слов в тексте использован конечный автомат с таблицей состояний, приведенной в табл. 1.

ТАБЛИЦА 1. Состояния управляющего автомата

Текущее состояние	Текущий символ	Новое состояние	Действие
1	Буква	2	Добавить символ в FIFO-память
1	Служебный символ	1	Пропустить символ
1	Разделитель	1	Пропустить символ
2	Буква	2	Добавить символ в FIFO-память
2	Служебный символ	2	Добавить символ в FIFO-память
2	Разделитель	1	Обработав слово, очистить FIFO-память

Автомат содержит FIFO-память для накопления текущего слова. Состояние конечного автомата определяет наличие в памяти накопленной части слова (1 - отсутствует, 2 - присутствует). К служебным символам отнесены апостроф и дефис, к разделителям - остальные символы, кроме букв. Состояние 1 — начальное, 1 и 2 - конечные. Если после выполнения конечного автомата его память не пуста, обрабатывается слово, расположенное в FIFO-памяти.

Комплекс включает более ста морфологических словарей общим объемом 70 МБ, охватывающих слова по частям речи и словоформам. На этапе инициализации словари распределяются по узлам кластера для балансировки вычислительной нагрузки. Для организации поиска словоформы в словаре использованы индексы в виде бинарного дерева поиска, особенность реализации которого - словоформы как атрибут поиска. Индексы хранятся в оперативной памяти, обеспечивая существенное ускорение обработки. Размер индексов не превышает половины объема словарей, а их распределение по узлам кластера существенно снижает объем индексов в оперативной памяти. Среднее количество слов обрабатываемых за 1 сек алгоритмом поиска, на ПК Pentium IV 1.6GHz составляет 143 слова, на кластере из 16 процессоров (2.66GHz) это количество достигает 1700 слов (см. рис. 2).

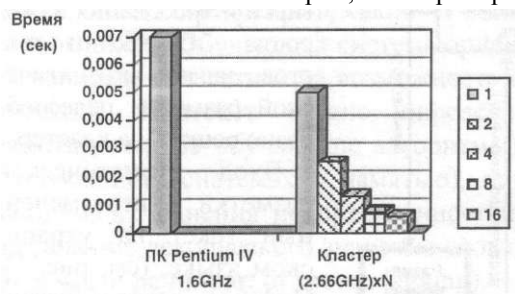

РИС. 2. Зависимость времени разметки одного слова от числа процессоров

Распараллеливание организовано в SPMD-архитектуре, согласно которой на всех узлах кластера выполняется одна и та же программа. Ее роль

в распределенном вычислении определяется рангом узла кластера (управляющий, управляемый) и задается номером, получаемым программой на старте.

Каждый словарь образуется комбинацией части речи с ее словоформами. Далее в таблицах 2, 3, 4 приведены возможные словоформы частей речи. Здесь использованы условные обозначения для словоформ: род (т - мужской, f- женский, п - средний); число (p - единственное, s - множественное, t - pluralia tantum для существительных только множественного числа, d - двойственное); склонение (п - именительный, g - родительный, d - дательный, a - винительный, i - творительный, l - предложный, V - призывной); степень сравнения (1 - высшая, 2 - наивысшая); вид (f - совершенный, c - несовершенный, d - двувидовая форма); наклонение (a - изъявительное, m - повелительное, j - сослагательное), время (p - настоящее, t - прошедшее, u - будущее); лицо (1 - первое, 2 - второе, 3 - третье); состояние (a - активное, v - пассивное). Дефис в ячейке таблицы указывает на неопределенную форму соответствующей части речи.

ТАБЛИЦА 2

Словоформа	Существительное	Прилагательное	Числительное
Род	m, f, n, -	m, f, n, -	m, f, n, -
Число	s, p, t, d	s, p, –	s, p, -
Склонение	n, g, d, a, i, l, v	n, g, d, a, i, l, v	n, g, d, a, i, l, -
Степень сравнения		1, 2	

ТАБЛИЦА 3

Словоформа	Порядковое числительное и местоимение	Глагол	Инфинитив
Род	m, f, n, -	m, f, n, -	
Число	n, g, d, a, i, l, -		
Вид		r, c	r, c, d
Наклонение		a, j, m, z	
Время		p, t, u, c	
Лицо		1, 2, 3, -	

ТАБЛИЦА 4

Словоформа	Причастие	Деепричастие	Наречие
Род	m, f, n, -	m, f, n, -	
Число	s, p, -		
Склонение	n, g, d, a, i, l, -		
Степень сравнения			1, 2
Вид	r, c	r, c	
Время	p, t		
Состояние	A, v		

Кроме вышеуказанных, в состав комплекса включены отдельные словари для союзов, предлогов и междометий.

Заключение. Эксперименты показали, что задача автоматической разметки поддается распараллеливанию на кластере, что подтверждает эффективность поддержки Национального корпуса украинского языка.

Автономность словарей по видам словоформ позволяет использовать априорную информацию о тексте (скажем, жанр и стиль текста) для балансировки вычислительной нагрузки на основании статистической информации о частотах вхождения видов словоформ в тексты разных видов и жанров. Именно в этом легко усмотреть дополнительную цель реализованного обучающего алгоритма.

О.М. Демська-Кульчицька, В.Р. Семеренко, Р.А. Ющенко

МЕТОДИ АВТОМАТИЧНОЇ РОЗМІТКИ ТЕКСТІВ НАЦІОНАЛЬНОГО КОРПУСУ МОВИ

Методи автоматичної розмітки розглядають у контексті дослідження природно мовних текстів, побудови та підтримки Національного корпусу української мови. Представлений огляд основних підходів. Наведено схему розпаралелювання одного з методів автоматичної розмітки на кластері.

O.M. Demska-Kulchitska, V.R. Semerenko, R.A.Yushchenko

METHODS OF AUTOMATIC NATURAL TEXT TAGGING FOR NATIONAL CORPORA

Methods for automatic tagging are presented in context of natural language processing research, for development and support of national Ukrainian corpora. Description to basic approaches is given. Parallel scheme for automatic tagging on cluster is presented.

1. *Демська-Кульчицька О.М.* Базові поняття корпусної лінгвістики // Українська мова. – 2003. – № 1. – С. 40–46.
2. *Демская-Кульчицкая О.М.* Корпус текстов украинской периодики // Исследование славянских языков в русле традиций сравнительно-исторического и сопоставительного языкознания. – М.: Изд-во Московского университета, 2001. – С. 26–38.
3. *Brill E.* A Simple Rule-Based Part of Speech Tagger // In Proceedings of the DARPA Speech and Natural Language Workshop. – San Mateo, California: Morgan Kauffman, 1997.– P. 112–116.
4. *Brill E.* Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of speech Tagging // Computational Linguistics. – 1995. – Vol. 21. – N 4. – P. 122–132.
5. *Михайлов М.* Параллельные корпуса художественных текстов. Академическая диссертация. – Тампере: Университет Тампере, 2003. – 162 с.
6. *Gale W.A., Church K.W., Yarowsky D.* A method for disambiguating word senses in a large corpus // Computers and the Humanities. – 1993. – Vol. 26. – N 1. – P. 415–439.
7. *Merialdo B.* Tagging English text with a probabilistic model // Computational Linguistics. – 1994. – Vol. 20. – N 2. – P. 155–172.

Получено 13.12.2004

Об авторах:

Демская-Кулъ чицкая Орыся Маръяновна,
кандидат филологических наук, заместитель директора Института украинского языка НАН Украины,

Семеренко Владислав Русланович,
студент Национального технического университета «КПИ»,

Ющенко Руслан Андреевич,
аспирант Института кибернетики им. В.М. Глушкова НАН Украины.