ПЕДАГОГИЧЕСКАЯ ИНФОРМАТИКА

ОСОБЕННОСТИ ПАРАЛЛЕЛЬНОЙ ОБРАБОТКИ РУССКОЯЗЫЧНОГО КОНТЕНТА С ИСПОЛЬЗОВАНИЕМ БАЗОВЫХ ХАРАКТЕРИСТИК ОБЪЕКТНО ОРИЕНТИРОВАННЫХ ЯЗЫКОВ ВЫСОКОГО УРОВНЯ

Д.Е. Кошкин

Московский государственный технический университет
радиотехники, электроники и автоматики
Проспект Вернадского, 78, Москва, Россия, 11945

А.К. Скуратов

Государственный научно-исследовательский институт
информационных технологий и телекоммуникаций
Брюсов переулок, 21, строение 2, Москва, Россия, 125009

В статье рассматриваются общие черты, прослеживаемые в русском языке и языках программирования высокого уровня и предлагается подход, выявляющий связи между словами в предложении и предложениями в рамках текста. Рассматриваемый вопрос может быть интересен как с точки зрения автоматического семантического анализа текстов, так и для систем автоматического перевода.

Ключевые слова: распределенная обработка текстов, естественный язык, языки программирования высокого уровня, семантическая единица.

Русский язык имеет официальный рабочий статус во множестве международных организаций (ООН, ОБСЕ, ШОС, ИСО) и присутствует в пятерке самых переводимых языков мира [1], что делает его достаточно важным для коммуникации и международного сотрудничества. В процессе международного взаимодействия создается множество документов, хранящих важную информацию, требующую обработки. Однако по сравнению с английским общее распространение русского языка в мире меньше. Для поддержки, укрепления и расширения сферы распространения русского языка была утверждена федеральная целевая программа «Русский язык» на 2011—1015 гг. [2]. Программа предусматривает разработку и апробацию справочно-информационных ресурсов в сфере русского языка, что может

быть рассмотрено и как исследования в области обработки текстов на русском языке и извлечения информации из них для дальнейшего использования результатов. Например, использование результатов исследования в системах автоматического перевода позволит повысить качество перевода с русского языка, использование их в информационно-поисковых системах приведет к повышению качества поисковой выдачи; результаты могут использоваться также в системе обучения русскому языку как носителей языка, так и учащихся из других государств.

Для автоматического создания семантических сетей из текстовых материалов на естественном языке необходимо формализовать структуру языка и обеспечить реализуемость этой формализации хотя бы на одном языке программирования. Рассмотрение текстов на естественном языке в свете максимально близких к ним языкам программирования позволит извлекать семантические отношения не только из отдельных предложений, но и из абзацев, если предложения в них связаны одной мыслью, а следовательно, и из целых текстов без предварительной разбивки и обработки.

На современном этапе развития теории интеллектуальной обработки данных алгоритмы, основанные на машинном обучении, адекватно определяют параметры ключевых объектов в предложении и позволяют строить связи внутри предложений [3; 4]. Однако построение связей между соседними предложениями часто вызывает затруднения. В решении этой задачи могут помочь модели, описывающее связи между предложениями через доступные характеристики слов — число, род, падеж, и т.д. Представим каждое слово объектом-наследником базового класса — части речи. У таких объектов присутствуют наследуемые параметры, влияющие на свойства самих объектов. Наборы параметров, характерные для определенной части речи, можно считать аналогами параметров классов в объектно ориентированном программировании (ООП). Попробуем найти и иные параллели между основными понятиями ООП и семантикой, грамматикой, морфологией русского языка.

В ООП используются понятия «класс», «объект», «абстракция», «инкапсуляция», «наследование» и «полиморфизм» [5]. Класс является моделью еще не существующей сущности (объекта), описанной на языке программирования. Фактически этот класс описывает общее устройство объекта, являясь своего рода его чертежом. Обычно классы разрабатывают таким образом, чтобы их объекты соответствовали объектам предметной области [6]. В русском языке аналогом класса является часть речи, которая содержит в себе набор необходимых параметров: число, род, падеж и др.

В табл. 1 приведены два примера объявления классов на языке программирования C++ и (для сравнения) части речи, оформленные по тем же правилам. В первом примере показывается, что объявление простого класса и его переменных похоже на то, какие характеристики хранят в себе все слова, принадлежащие к одной части речи. Во втором примере показывается сходство объявления класса на основе уже существующих классов и причастия — части речи, производной от глагола и прилагательного.

Такое сравнение показывает, что части речи можно воспринимать как классы в языках программирования.

Таблица 1

Сравнение объявления классов в C++ и частей речи в русском языке

	Язык программирования C++	Русский язык

1	class MyClass {
	private:
	uint dd;
	uint mm;
	uint yy;
	public:
	uint& Day ();
	uint& Month ();
	uint& Year ();
	}
		Имя существительное{
		private:
		собственное или нарицательное;
		одушевленное или неодушевленное;
		тип склонения;
		public:
		падеж;
		число
		род;
		}

2	class MyClass: public ParentClass1,
	public ParentClass2
	{
	public:
	MyClass();
	~MyClass();
	int& ClassMember();
	private:
	int classmember;
	}
		Причастие: Глагол, Прилагательное {
		вид;
		действительное
		возвратное или нет
		переходное или нет
		страдательное
		краткое или полное
		время;
		число;
		род (только в ед. ч.);
		падеж;
		}

В дальнейшем под объектом будем понимать экземпляр класса — сущность в адресном пространстве вычислительной системы, появляющуюся при создании экземпляра класса или копирования прототипа. Исходя из этого определения объект-наследник класса обладает если не всеми, то большинством его характеристик, при этом слово является экземпляром класса «часть речи» и несет в себе признаки класса-родителя.

«Абстракция данных» — популярная и, как полагают, не всегда верно определяемая техника программирования [5]. Фундаментальная идея состоит в разделении несущественных деталей реализации подпрограммы и характеристик, существенных для корректного ее использования [5]. Для корректного использования части речи в словосочетаниях и предложениях большую роль играет несколько характеристик — род, число, падеж. Эти характеристики в большей мере задают формообразующие морфемы слова, в частности окончания. Например, несущественным для существительного может быть тип склонения, а существенным — род.

Свойством системы, позволяющим объединить данные и методы, работающие с ними, в классе и скрыть детали реализации от пользователя, является инкапсуляция. В естественном языке под инкапсуляцией можно понимать правила спряжения и склонения частей речи с присвоением соответствующего окончания или изменение параметров рода, числа, падежа, вида и других параметров.

Свойство системы, позволяющее описать новый класс на основе уже существующего, с частично или полностью заимствующейся функциональностью называют наследованием [5]. Класс, от которого производится наследование, называется базовым, родительским или суперклассом, а новый класс является потомком,

наследником или производным классом. Как ранее отмечено, каждое слово причисляется к определенной части речи, от которой наследует список характеристик, заложенный в классе-родителе. Свойство системы использовать объекты с одинаковым интерфейсом без информации о типе и внутренней структуре объекта определяют как ее полиморфизм [5]. Использование в предложениях слов как базовых блоков также вовлекает во взаимодействие части речи. Так, если в предложении одно прилагательное заменяется другим с равными характеристиками, строй предложения не изменится, хотя может измениться смысл.

Исходя из примеров, аналогий и параллелей, приведенных выше, напрашивается вывод о том, что текст на практически любом естественном языке должен иметь общие черты с исходным кодом компьютерной программы. Следовательно, должны существовать правила семантики «языка программирования», которые могут быть формализованы и представлены в виде, пригодном для машинной обработки. При этом нельзя исключать погрешности, возникающие при формализации, которые устраняются при коррекции на экспериментах и при накоплении статистической и фактической информации. В построении связей между предложениями будет видна «замена переменных», выглядящая объявлением новой переменной, и дальнейшим ее изменением. Например, в научных текстах объект, о котором говорится впервые, в предложении обычно обозначается существительным в именительном или дательном падеже. Во втором предложении, если речь идет о том же объекте, — местоимением, но с сохранением числа и рода. Также некоторые виды частей речи можно автоматически изменять, руководствуясь принципами перегрузки методов и переопределения классов. Притяжательное местоимение «его» может заменяться на словосочетание «принадлежит ему» или «он владеет», что автоматически указывает на объект, упомянутый ранее, и отношения между текущим и предыдущим объектами. По этим аналогиям можно строить связи между двумя и более предложениями, основываясь на схожих параметрах слов и выражая их через граф, с нагруженными ребрами. При этом сам граф — представление семантических отношений между предложениями, а нагрузки ребер графа выражают отношения между вершинами-объектами.

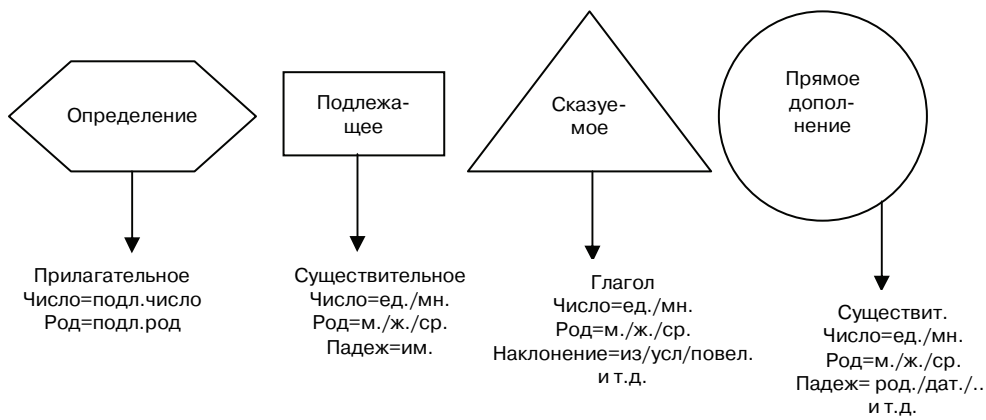

Рис. 1. Пример рассмотрения параметров объектов в рамках предложения

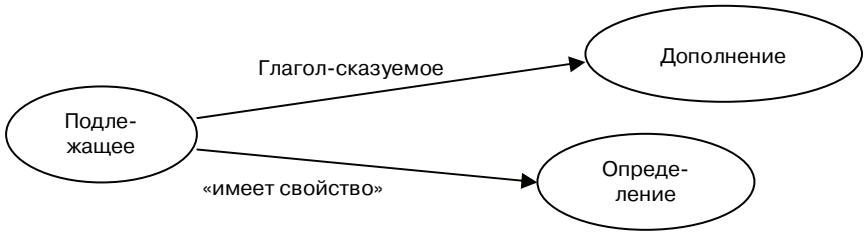

Рис. 2. Две ветви графа, отражающие связи из рис. 1

Особенностью русскоязычных текстов является их модульность, где каждый модуль является предложением. В общем случае обработка каждого предложения может вестись независимо от всего текста, а значит, параллельно на множестве одновременно работающих вычислителей. Параллельную обработку обеспечивают две технологии: высокопроизводительные вычислительные кластеры и GRID-системы [7; 8]. Использование глобальных распределенных вычислительных сетей типа GRID-систем для обработки отдельных предложений нецелесообразно ввиду малого времени обработки последних. Для GRID-систем обработка одного текста на одном вычислителе является наиболее рациональным использованием ресурсов по соотношению время обработки/время передачи данных.

В высокопроизводительных вычислительных кластерах, наоборот, при общей памяти и общем КЭШе время передачи данных и время обработки могут составлять сверхмалые величины, так как скорость межузлового взаимодействия значительно выше, чем у узлов GRID-системы. При разделении текстов на предложения представляют интерес шаги, которые предпринимаются высокопроизводительными вычислительными кластерами для обработки текстов и поиска связей между словами в предложении и самими предложениями.

Первым шагом, разложением текста на отдельные предложения и формированием из них соответствующего для обработки материала занимается главный узел кластера. Он получает текст из хранилища, заносит его данные в общую базу, разделяет на предложения и дает узлам команду начать обработку структурных единиц предложения.

Вторым шагом является обработка и анализ структурных единиц. Под обработкой каждой структурной единицы здесь понимается «тегирование» — присвоение слову определенных параметров, которые в текущем случае подразумевают часть речи, число, род, падеж и др., которые могли бы быть полезными в процессе нахождения связи между предложениями. В процессе анализа для каждого слова могут быть поставлены в соответствие несколько таких групп параметров. Получающееся множество будет частично упорядочиваться при связывании слов, частично предложений, хотя случаи неполного определения соответствия не исключены. В таком случае в зависимости от используемого алгоритма или модуля тегирования снятие неопределенности может явиться прерогативой эксперта. Второй вариант менее предпочтителен ввиду перехода от автоматической к авто-

матизированной обработке. Однако в особо сложных или противоречивых случаях при первых запусках системы использование человека-специалиста может принести пользу.

Третий шаг финальный: сервер совмещает предложения в единый массив данных, который обрабатывается на одном вычислителе. На этом шаге используется приведенная выше логика преобразования на основе сходных параметров у разных частей речи, сообщающих об одном и том же объекте. Поэтому на граф связей наносятся связи не только между словами одного предложения, но и связи между предложениями. При использовании определенного типа отношений для создания графа сам граф будет сходен по структуре с семантическими сетями.

Общая практическая значимость подхода заключается в возможности представлять любые текстовые материалы в виде базы знаний, повышая пертинентность ответа поисковой системы в разы и ускоряя поиск актуальной информации. Логически построенная структура одного текста позволит объединить идентичными ветвями с другими текстами и расширять хранимую базу. В общем случае автоматическая обработка текстов приблизит переход к семантическим порталам и, возможно, к семантическому Интернету, избавив пользователей от необходимости перечитывать десятки страниц в Сети ради поиска одного факта.

В статье рассмотрены черты языков программирования высокого уровня, прослеживаемые в семантике русского языка. Предлагается подход, выявляющий связи между словами в предложении и предложениями в рамках текста. Рассмотрено техническое обеспечение этого подхода и созданы основные предпосылки для разработки алгоритмов распределенной обработки текстов. Представленный подход позволит перевести задачу автоматизированного построения онтологий предметных областей в задачу автоматического их построения, приближая переход к семантическому Интернету из далекого во вполне обозримое будущее.

ЛИТЕРАТУРА

[1] Английский, французский, немецкий и русский языки — самые переводимые в мире // CyberSecurity.ru — 2012. — 19 апреля. — URL: http://www.cybersecurity.ru/prognoz/149230.html

[2] Постановление Правительства РФ от 20.06.2011 № 492 (ред. от 02.04.2012) «О федеральной целевой программе „Русский язык“ на 2011—2015 годы». — URL: http://fcp.economy.gov.ru/cgi-bin/cis/fcp.cgi/Fcp/ViewFcp/View/2015/306

[3] *Пак А.* Определение части речи слов в русском тексте (POS-tagging) на Python 3 // Тематические медиа. — 2011 // URL: http://habrahabr.ru/post/125988/

[4] *Пак А.* Парсим русский язык // Тематические медиа. — 2012. — URL: http://habrahabr.ru/post/148124/

[5] *Страуструп Б.* Язык программирования C++, Специальное издание / Пер. с англ. — СПб.: БИНОМ, 2001.

[6] *Буч Г.* Объектно-ориентированный анализ и проектирование с примерами приложений на C++ = Object-Oriented Analysis and Design with Applications / Пер. И. Романовский, Ф. Андреев. — 2-е изд. — М., СПб.: Бином, 1998.

[7] *Дробнов С.Е., Кошкин Д.Е.* Расчет оптимального количества вычислителей GRID-системы // Современные информационные технологии в управлении и образовании: Сб. научн. тр. В 3-х ч. — М.: Восход, 2012. — Ч. 1.

[8] *Дробнов С.Е., Кошкин Д.Е.* Анализ ускорения обучения нейронных сетей при применении GRID-систем // Материалы IV Всероссийской конференции студентов, аспирантов и молодых ученых «Искусственный интеллект: философия, методология инновации» (Москва, МИРЭА, 10—12 ноября 2010 г.) / Под ред. Д.И. Дубровского и Е.А. Никитиной. — М.: Радио и Связь, 2010. — Часть I. — 168 с.

[9] *Скуратов А.К., Илиева С.Ю., Пашовкина Н.А.* Информационное обеспечение международного сотрудничества // Вестник Российского университета дружбы народов. Серия «Информатизация образования». — 2012. — № 4. — С. 5—10.

[10] *Sigalov A., Skuratov A.* Educational Portals and Open Educational Resources in the Russian Federation // UNESCO Institute for Information Technologies in Education. — Moscow, 2012.

LITERATURA

[1] Anglijskij, francuzskij, nemeckij i russkij jazyki — samye perevodimye v mire // CyberSecurity.ru — 2012. — 19 aprelja. — URL: http://www.cybersecurity.ru/prognoz/149230.html

[2] Postanovlenie Pravitel'stva RF ot 20.06.2011 № 492 (red. ot 02.04.2012) «O federal'noj celevoj programme „Russkij jazyk“ na 2011—2015 gody». — URL: http://fcp.economy.gov.ru/cgi-bin/cis/fcp.cgi/Fcp/ViewFcp/View/2015/306

[3] *Pak A.* Opredelenie chasti rechi slov v russkom tekste (POS-tagging) na Python 3 // Tematicheskie media. — 2011. — URL: http://habrahabr.ru/post/125988/

[4] *Pak A.* Parsim russkij jazyk // Tematicheskie media. — 2012. — URL: http://habrahabr.ru/post/148124/

[5] *Straustrup B.* Jazyk programmirovanija C++, Special'noe izdanie / Per. s angl. — SPb.: BINOM, 2001.

[6] *Buch G.* Ob#ektno-orientirovannyj analiz i proektirovanie s primerami prilozhenij na S++ = Object-Oriented Analysis and Design with Applications / Per. I. Romanovskij, F. Andreev. — 2-e izd. — M., SPb.: Binom, 1998.

[7] *Drobnov S.E., Koshkin D.E.* Raschet optimal'nogo kolichestva vychislitelej GRID-sistemy // Sovremennye informacionnye tehnologii v upravlenii i obrazovanii: Sb. nauchn. tr. V 3 ch. — M.: Voshod, 2012. — Ch. 1.

[8] *Drobnov S.E., Koshkin D.E.* Analiz uskorenija obuchenija nejronnyh setej pri primenenii GRID-sistem // Materialy IV Vserossijskoj konferencii studentov, aspirantov i molodyh uchenyh «Iskusstvennyj intellekt: filosofija, metodologija innovacii» (Moskva, MIRJeA, 10—12 nojabrja 2010 g.) / Pod red. D.I. Dubrovskogo i E.A. Nikitinoj. — M.: Radio i Svjaz', 2010. — Chast' I.

[9] *Skuratov A.K., Ilieva S.Ju., Pashovkina N.A.* Informacionnoe obespechenie mezhdunarodnogo sotrudnichestva // Vestnik Rossijskogo universiteta druzhby narodov. Serija «Informatizacija obrazovanija». — 2012. — № 4. — S. 5—10.

[10] *Sigalov A., Skuratov A*. Educational Portals and Open Educational Resources in the Russian Federation // UNESCO Institute for Information Technologies in Education. — Moscow, 2012.

FEATURES OF PARALLEL PROCESSING OF RUSSIAN LANGUAGE CONTENT USING THE BASIC CHARACTERISTICS OF OBJECT-ORIENTED HIGH-LEVEL PROGRAMMING LANGUAGES

D.E. Koshkin

Moscow state technical university
radio engineering, electronics and automatic equipment
Prospekt Vernadskogo, 78, Moscow, Russia, 11945

A.K. Skuratov

State research institute
information technologies and telecommunications
Brjusov pereulok, 21, structure 2, Moscow, Russia, 125009

The article discusses the general features which are traceable in the Russian language and high-level programming languages, and offers an approach to identify the relationship between words in a sentence and proposals within the text. A matter may be interesting in terms of automatic semantic analysis of texts and for the systems of automatic translation.

Key words: distributed processing, natural language, high-level programming languages, semantic unit.