УДК 004.048:519.767.6

ЛИНГВИСТИЧЕСКИЕ ОСНОВЫ АЛГОРИТМОВ КОМПЬЮТЕРНОЙ ОБРАБОТКИ ТЕКСТОВ НА ОСНОВЕ СИСТЕМ С ПРЕДОПРЕДЕЛЕННОЙ СЕМАНТИКОЙ

Ванюлин А.Н., Алексеева Н.Р., Мочалова Т.А.

ФГБОУ ВО «Чувашский государственный университет им. И.Н. Ульянова», Чебоксары, e-mail: van-u-lin@yandex.ru

В статье рассматриваются вопросы создания систем автоматизированной обработки текстов. Отмечено, что имеющиеся на рынке системы практически полностью удовлетворяют потребности пользователей. Показано, что при программной реализации в зависимости от назначения системы используются различные современные методы и технологии, с общим названием «технологии искусственного интеллекта». Однако все эти технологии совершенно не учитывают особенности текстов на естественном языке и не имеют общего, характерного для человеческого интеллекта признака – универсальности, т.е. отсутствует возможность их применения для решения любой задачи. Для разрешения данного вопроса предлагается использование алгоритмов на основе систем с предопределенной семантикой. Описаны основные понятия подобных систем и базовые алгоритмы их реализации. В основе алгоритмов лежат данные лингвистики о структуре предложений на естественном языке. В результате экспериментов с программным прототипом показано, что после обучения создаваемая система практически автоматически учитывает основные особенности естественной речи и позволяет с высокой степенью надежности идентифицировать тексты, различного содержания. Показано, что заложенные в систему алгоритмы обладают высокой степенью универсальности и могут быть использованы для создания приложений различного назначения.

Ключевые слова: тексты на естественном языке, лингвистический процессор, искусственный интеллект, алгоритмы, предопределенная семантика

LINGUISTIC FOUNDATIONS OF ALGORITHMS IN COMPUTER PROCESSING OF TEXTS ON THE BASIS OF SYSTEMS WITH A PREDEFINED SEMANTICS

Vanyulin A.N., Alekseeva N.R., Mochalova T.A.

Federal State Educational Budget Institution of Higher Education «The Ulianov Chuvash State University», Cheboksary, e-mail: van-u-lin@yandex.ru

The article considers the problems of developing natural language processing systems. The authors outline that current systems on the market almost satisfy users' needs. These systems are developed by means of various modern methods and technologies, depending on the purpose of the system, with the general name of artificial intelligence technologies. However, all these technologies do not consider the peculiarities of natural language texts and do not have a common characteristic of human intelligence – flexibility, i.e. they provide no possibility to solve any problem. To resolve this issue, the authors suggest using algorithms based on systems with predefined semantics. The article describes basic concepts of such systems and basic algorithms for their implementation. At the core of algorithms is linguistic data about the structure of sentences in a natural language. As a result of experiments with a prototype program product the authors demonstrate that the system after a learning process almost automatically considers basic peculiarities of natural speech and identifies texts with different content with a high degree of reliability. The article shows that the algorithms embedded in the system have a high degree of flexibility and can be used to create applications for various purposes.

Keyword: natural language texts, linguistic processor, artificial intelligence, algorithms, predefined semantics

Попытки использовать вычислительную технику для автоматизированной обработки текстовой информации начались сразу же после появления первых ЭВМ. К настоящему времени сложился ряд устойчивых направлений применений информационных технологий для обработки текстов – от классических текстовых редакторов до систем машинного перевода.

Разрабатываемые технологии стали основой для совершенно нового научного направления – компьютерной лингвистики (КЛ). В общем случае задачи КЛ могут быть сформулированы как разработка методов и средств построения лингвистических процессоров (ЛП) для различных прикладных задач по автоматической обработке текстов на естественном языке (ЕЯ).

Можно выделить следующие основные направления, решаемые методами КЛ [1, 2]: машинный перевод (Machine Translation); информационный поиск (Information Retrieval); задачи индексирования, реферирования, аннотирования и классификации документов; TextMining (извлечение информации из текстов) [3]; DataMining (интеллектуальный анализ данных) [4]; системы поддержки диалога на ЕЯ.

При этом неявно предполагается, что все виды обработки текстовой информации очень близки к интеллектуальной деятельности человека, а потому при создании прило-

жений в КЛ используются все методы и технологии искусственного интеллекта (ИИ).

Однако словосочетание «искусственный интеллект» подразумевает использование технологий, полностью имитирующих механизмы интеллектуальной деятельности человека. Основным признаком такой деятельности является то, что данный механизм позволяет не только использовать уже известные алгоритмы, но и самостоятельно вырабатывать новые алгоритмы для решения новых, еще не решаемых системой задач.

Согласно этому признаку ни одна из так называемых систем ИИ не может называться таковой, поскольку использует уже готовые, разработанные человеком алгоритмы.

Целью настоящей работы является разработка универсальных алгоритмов, позволяющих реализовать задачи обработки неструктурированной текстовой информации. Основные требования к алгоритмам заключаются в следующем:

– алгоритмы должны учитывать все особенности текстов на ЕЯ;

– алгоритмы должны быть достаточно универсальными, т.е. применимыми для широкого класса задач автоматизированной обработки текстов.

Для реализации лингвистических процессоров различного назначения применяются достаточно сложные алгоритмы, использующие все современные методы обработки больших данных, математической статистики и технологии искусственного интеллекта.

Например, для систем автоматической классификации текстов применяются как стандартные методы поиска в сочетании с классическими статистическими методами анализа, так и такие современные методы, как нейронные сети, деревья решений, генетические алгоритмы и т.д. [5]. Достигаемая точность классификации может составлять свыше 90%. При этом основой всех алгоритмов является использование наборов ключевых слов и частотности их применения в различных предметных областях. Очевидно, что ни о какой интеллектуальной деятельности в виде распознавания смысла текста говорить не приходится – просто производится механическая классификация текстов с использованием современных методов машинного обучения.

В системах машинного перевода основными алгоритмами являются:

– пословный перевод;

– использование алгоритмов построения структуры переводимого предложения с последующим заполнением такой же структуры словами-аналогами другого языка. Данный алгоритм хорошо зарекомендовал себя при работе с текстами родственных языков (русский – украинский, сербский – хорватский и т.д.);

– алгоритмы с использованием промежуточного языка, на котором кодируется смысл переводимых фраз. Данная технология позволяет успешно работать с текстами языков, структуры которых существенно отличаются.

Все перечисленные типы алгоритмов постоянно совершенствуются в основном за счет использования статистики переводных пар слов и словосочетаний.

На практике автоматические переводчики используются в трех случаях:

– перевод художественной литературы. В этом случае переведенный системой текст подвергается практически стопроцентной переработке, и в лучших случаях результат перевода становится отдельным художественным произведением;

– перевод научной литературы. Если речь идет о публикации в иностранных научных журналах, то получаемый системой перевод также подвергается стопроцентной переработке. Связано это с тем, что в каждой предметной области используется своя специфическая терминология и методы подачи материала;

– извлечение информации из иноязычных текстов. В данном случае требования к качеству переведенного текста практически отсутствуют – достаточно лишь того, чтобы полученный текст достаточно адекватно передавал смысл исходного сообщения.

Практика использования имеющихся систем показывает, что они практически полностью удовлетворяют потребности пользователей.

Для систем поддержки диалога на естественном языке конкретная алгоритмическая реализация является, конечно, коммерческой тайной. Поэтому главным источником информации по данному вопросу являются интернет-публикации. Например, согласно [6], основными технологиями создания систем поддержки диалога являются: нейронные сети, регулярные выражения и машинное обучение.

Типичными такими системами являются Алиса от Яндекса, Alexa от Amazon, Google Assistant от Google, Siri от Apple и Cortana от Microsoft. Данные системы могут вести достаточно осмысленный диалог, определяют интенты (намерения) пользователей и могут выполнять какие-то команды. Часть из них даже проходит тест Тьюринга.

Еще одной особенностью, характерной для всех групп ЛП является то, что кроме использования различных алгоритмиче-

ских решений для создания ЛП приходится привлекать большие группы специалистов из соответствующих предметных областей. Все это приводит к повышению как общей стоимости создаваемых программных продуктов, так и к необходимости начинать разработки практически с нуля при создании новых типов ЛП.

В то же время для решения всех этих типов задач возможны и другие подходы. Это, например, использование систем с предопределенной семантикой [7].

В основу таких систем положены следующие базовые понятия.

Понятие семы

Сема представляет собой элемент либо самой системы, либо внешнего мира, известный системе с момента начала ее функционирования. Список исходных сем составляется исходя из возможностей доступа системы к основным ресурсам компьютера. Это, например, файл, папка, диск, удалить, копировать и т.д. В этот список также включаются такие понятия внешнего мира, которые недоступны системе или их очень трудно объяснить с помощью других понятий. Это, например, запах, вкус, жидкость и т.п. Кроме того, в зависимости от назначения системы в этот же список можно добавить понятия, связанные с конкретной предметной областью. Это дает возможность ускорить будущее обучение системы.

Именно это предположение, изначально задающее определенное количество базовых семантических компонент, и обуславливает общее название подобных систем.

Понятие семантического образа

Семантический образ представляет собой упорядоченное множество значений базовых сем. Например, если при программной реализации определено 100 базовых сем и соответствующих словесных эквивалентов, то любой семантический образ будет представлять собой массив, состоящий из ста элементов числового вещественного типа. Данный массив будет представлять собой семантику каждого слова в виде некоторого семантического спектра. В исходном состоянии у необученной системы будет набор только базовых понятий. Для них семантический спектр будет состоять только из одного элемента, номер которого определен в списке базовых понятий. Предполагается, что в процессе обучения системы семантика новых слов и понятий будет формироваться путем смешения спектров базовых понятий в определенных пропорциях. Примеры визуального представления спектров приведены в табл. 2.

Лингвистические основы системы

Основные особенности текстов на ЕЯ рассмотрены в [8]. Для учета этих особенностей используется следующее положение:

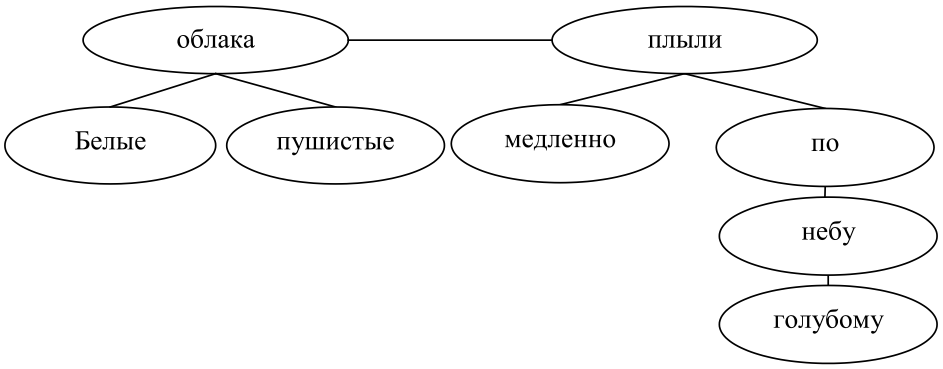
– в лингвистике принято представление отдельных фраз и предложений в виде некоторой структуры [7]. Например, предложение:

«Белые, пушистые облака
медленно плыли по голубому небу» (1)

может быть представлено следующей структурой (рисунок).

Профессиональные лингвисты едва ли используют понятия теории графов в своей повседневной деятельности, но такое представление очень удобно и наглядно показывает взаимосвязи слов в предложении. При этом смысл взаимосвязей сводится к следующему:

– основной смысл предложения содержится в словах, находящихся на самом верхнем уровне дерева. Слова же, находящиеся на более низких уровнях, просто уточняют смысл соответствующих слов верхнего уровня.



Структурная схема предложения

С точки же зрения информатики это просто представление структуры предложения в виде дерева, опираясь на которую необходимо рассчитать его общую семантику.

Для этого предлагается использовать следующий алгоритм:

1. Из базы данных загружается семантика всех слов, входящих в предложение, и размещается в узлах дерева.

2. Обработка дерева производится слева направо и начинается с терминальных узлов дерева.

При этом расчет семантик для каждой пары узлов нижнего и верхнего уровня предлагается производить по следующей формуле:

$$s_0 = (s_1 + s_2/2)/2, \quad (2)$$

где s_0 – обобщенная семантика связанной пары узлов;

s_1 – семантика узла верхнего уровня;

s_2 – семантика узла нижнего уровня.

Логическое обоснование применения формулы (2) заключается в том, что слова верхнего уровня несут более значимую информацию, а слова более низкого уровня несут лишь уточняющую информацию. Поэтому вклад слов верхнего уровня в общую семантику должен быть больше. Для слов самого верхнего уровня дерева также применяется формула (2), но здесь в качестве семантики верхнего уровня используется семантика контекста, которая может быть и нулевой.

Таким образом, общая семантика фразы вычисляется по формуле

$$s = \sum_{i=1}^{n} k_i s_i, \quad (3)$$

где n – количество слов в предложении;

s_i – семантика i-го слова;

k_i – структурный коэффициент i-го слова, значение которого зависит от конфигурации дерева и отражает вклад семантики данного слова в общую семантику фразы.

Например, для фразы (1) со структурой, приведенной на рисунке, вычисления будут выглядеть следующим образом:

$$s = ((s_{\text{контекст}} + (s_{\text{облака}} + s_{\text{белые}}/2)/2)/2 + s_{\text{пушистые}}/2)/2 + ((s_{\text{плыли}} + s_{\text{медленно}}/2)/2 + (s_{\text{по}} + (s_{\text{небу}} + s_{\text{голубому}}/2)/2)/2)/2. \quad (4)$$

Тогда соответствующие значения структурных коэффициентов будут иметь значения, приведенные в табл. 1.

Нетрудно заметить, что чем ниже слово в дереве, тем меньше вклад его семантики в общую семантику предложения. При этом лингвисты всегда могут привести массу примеров, в которых основной смысл предложения может оказаться сосредоточен именно в словах низкого уровня. Но это возражение снимается тем, что при программной реализации система сама строит необходимые деревья, вид которых непривычен лингвисту, но зато их структура и рассчитываемая с ее помощью, семантика более полно соответствует смыслу предложений.

Благодаря описанной схеме обработки дерева автоматически решается вопрос о зависимости смысла текста от порядка слов. В качестве примера в табл. 2 приведены семантически спектры исходных слов и итоговой семантики следующих фраз:

«Глубина два метра», (5)

«Глубина метра два». (6)

Различия в спектрах можно оценить не только визуально, но численно. Для этого можно использовать, например, стандартный коэффициент корреляции. В случае идентичных спектров его значение должно быть равно единице. В данном же случае его значение равно 0,65.

Кроме того, использование предлагаемого подхода позволяет учитывать и другие особенности текстов на естественном языке. В частности, многозначность слов (полисемию) и влияние контекста на смысл фразы.

Укажем основные направления, при реализации которых можно использовать разработанные алгоритмы.

Например, для систем классификации текстов достаточно сформировать семантические спектры текстов, соответствующие различным предметным областям. При этом имеются все основания предполагать, что обработка множества текстов по какой-то конкретной предметной области и последующее их обобщение приведет к получению специфических и статистически устойчивых видов спектров.

Таблица 1

Значения структурных коэффициентов семантик отдельных слов для фразы (1)

Слово	Контекст	Белые	пушистые	облака	медленно	плыли	по	голубому	небу
Структурный коэффициент	1/4	1/16	1/8	1/8	1/16	1/8	1/8	1/32	1/16

Таблица 2

Вид семантических спектров слов и фраз, составленных на их основе

Слово/текст	Семантический спектр
Глубина	
два	
метра	
Фраза (5)	
Фраза (6)	

Опираясь на имеющуюся базу данных спектров, процесс классификации может быть произведен по следующему алгоритму:

– формируется семантический спектр обрабатываемого текста;

– полученный спектр поочередно сопоставляется с базой данных спектров различных предметных областей. В результате выбирается та предметная область, спектр которой в наибольшей степени согласуется со спектром обрабатываемого текста.

Аналогичный алгоритм можно использовать и при создании командных систем для интеллектуального управления техническими устройствами [9].

В данном случае заранее формируется база данных семантики соответствующих команд. После получения конкретной команды формируется ее семантика и из базы данных выбирается команда с наиболее согласованным спектром.

Заключение

Разработанные к настоящему времени системы компьютерной обработки текстов на практике дают вполне удовлетворительные результаты. Однако они имеют следующие недостатки:

– для каждого вида приложения используются свои специфические алгоритмы;

– использование сторонних сервисов, например использование баз данных для проведения морфемного анализа, модулей синтаксического и семантического анализа;

– требуют привлечения специалистов из других предметных областей (чаще всего профессиональных лингвистов);

– отсутствие универсальности. Для каждого типа приложения приходится применять или специально разрабатывать совершенно разные технологии.

Предлагаемый подход устраняет практически все указанные недостатки, поскольку при создании приложений любого назначения используются одни и те же алгоритмы.

Список литературы

1. Николаев И.С., Митренина О.В., Ландо Т.М. Прикладная и компьютерная лингвистика: коллективная монография. М.: URSS, 2016. 320 с.

2. Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие. М.: МИЭМ, 2011. 272 с.

3. Горковенко Д.К. Применение методов text mining для классификации информации, распространяемой в социальных сетях // Молодой ученый. 2016. № 14 (118). С. 66–72. [Электронный ресурс]. URL: https://moluch.ru/archive/118/32878/ (дата обращения: 20.02.2020).

4. Daniel Jurafsky, James H. Martin. Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. 2019. 613 p. [Электронный ресурс]. URL: https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf (дата обращения: 20.02.2020).

5. Люгер Джордж Ф. Искусственный интеллект. Стратегии и методы решения сложных проблем, 4-е изд.: пер. с англ. М.: Издательский дом «Вильямс», 2003. 864 с.

6. Разговорный AI: как работают чат-боты и кто их делает. 2018. [Электронный ресурс]. URL: https://habr.com/ru/company/just_ai/blog/364149/ (дата обращения:15.09.2019).

7. Касаткин Л.Л., Клобуков Е.В., Крысин Л.П. Русский язык: учебник для студентов учреждений высшего профессионального образования / Под ред. Л.Л. Касаткина. 4-е изд., перераб. М.: Издательский центр «Академия», 2011. 780 с.

8. Ванюлин А.Н., Шабалина Т.А. Особенности текстов на естественном языке при их компьютерной обработке // Состояние и перспективы развития ИТ-образования: сб. докл. и научн. ст. Всероссийской научн.-практ. конф. Чебоксары: Изд-во Чуваш. ун-та, 2019. С. 377–381.

9. Ванюлин А.Н. Алгоритм распознавания смысла команд на основе систем с предопределенной семантикой // Состояние и перспективы развития ИТ-образования: сб. докл. и научн. ст. Всероссийской научн.-практ. конф. (посв. 50-летию Чувашского гос. ун-та им. И.Н. Ульянова). Чебоксары: Изд-во Чуваш. ун-та, 2018. С. 199–206.