СТАТЬИ

УДК 004.048:519.767.6

АЛГОРИТМЫ РЕАЛИЗАЦИИ СИСТЕМ С ПРЕДОПРЕДЕЛЕННОЙ СЕМАНТИКОЙ НА ОСНОВЕ КОНЦЕПЦИИ СЕМАНТИЧЕСКИХ ПОЛЕЙ

Ванюлин А.Н., Алексеева Н.Р.

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

Системы с предопределенной семантикой являются достаточно универсальными при создании приложений для автоматизированной обработки текстов различного назначения. Основой таких систем является набор заранее определенных элементарных семантических элементов – сем. В самом простом варианте семами могут быть отдельные символы клавиатуры. Комбинация сем представляет смысл обрабатываемого текста. Такие комбинации можно представить в виде некоторого спектра, уникальность которого гарантирует распознавание смысла как отдельных фраз, так и текста в целом. В настоящей работе для создания подобных систем предлагается другой вариант – на основе концепции семантических полей. Данный подход является несколько более обоснованным с точки зрения данных о механизмах мышления человеческого мозга. Рассмотрены алгоритмы построения подобных систем, описана процедура обучения, рассмотрены вопросы создания диалоговых систем на их основе. Показано, что при обработке текстов системой учитываются практически все особенности естественного языка. Так же как и при использовании семантических спектров, эксперименты с программным прототипом показывают, что при его работе автоматически учитываются основные особенности естественной речи и подтверждаются данные лингвистики о структуре предложений на естественном языке.

Ключевые слова: алгоритм, обработка текстов, предопределенная семантика, семантические поля, машинное обучение

ALGORITHMS FOR IMPLEMENTING SYSTEMS WITH PREDEFINED SEMANTICS BASED ON THE CONCEPT OF SEMANTIC FIELDS

Vanyulin A.N., Alekseeva N.R.

The Ulianov Chuvash State University, Cheboksary, e-mail: van-u-lin@yandex.ru

Systems with predefined semantics are quite versatile when creating applications for automated text processing for various purposes. The basis of such systems is a set of predefined elementary semantic elements – semes. In the simplest version, the semes can be individual keyboard characters. The combination of these represents the meaning of the text being processed. Such combinations can be represented in the form of a certain spectrum, the uniqueness of which guarantees the recognition of the meaning of both individual phrases and the text as a whole. In this paper, another option is proposed for the creation of such systems – based on the concept of semantic fields. This approach is somewhat more reasonable in terms of data on the mechanisms of thinking of the human brain. The algorithms of building such systems are considered, the training procedure is described, the issues of creating dialog systems based on them are considered. It is shown that when processing texts, the system takes into account almost all the features of the natural language. Just as with the use of semantic spectra, experiments with the program proto-type show that when it works, the main features of natural speech are automatically taken into account and linguistics data on the structure of sentences in natural language are confirmed.

Keywords: algorithm, text processing, predefined semantics, semantic fields, machine learning

В работах [1; 2] показано, что системы с предопределенной семантикой являются достаточно универсальными при создании приложений для автоматизированной обработки текстов различного назначения. Основой таких систем является набор заранее определенных элементарных семантических элементов – сем, комбинации которых формируют семантический спектр. В простейшем случае набором элементарных сем могут являться все символы клавиатуры.

Правило формирования спектра заключается в последовательном добавлении к итоговому спектру семантики очередного символа текста/слова. При этом начальное значение семы, номер которой совпадает с ASCII-кодом символа, равно единице, а остальные семы равны нулю. При добавлении последующих символов производится корректировка всех сем итогового спектра по формуле:

$$S_{ti} = (S_{ti} + S_{s1}) / 2, \quad (1)$$

где S_{ti} – i-я сема текста;

S_{si} – i-я сема очередного символа текста.

Полученный спектр характеризует смысл обрабатываемого текста. В качестве примера на рис. 1 представлен семантический спектр слова «привет». На виде спектра отражается действие формулы (1) – последний символ имеет максимальное значение, а первый – минимальное. Эта особенность спектра и позволяет учитывать порядок следования символов в тексте.

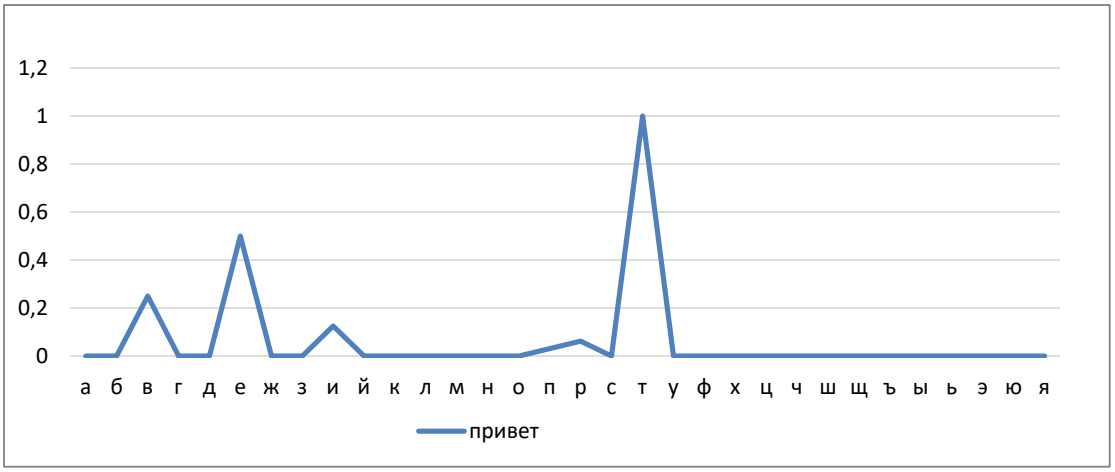


Рис. 1. Семантический спектр слова «привет»

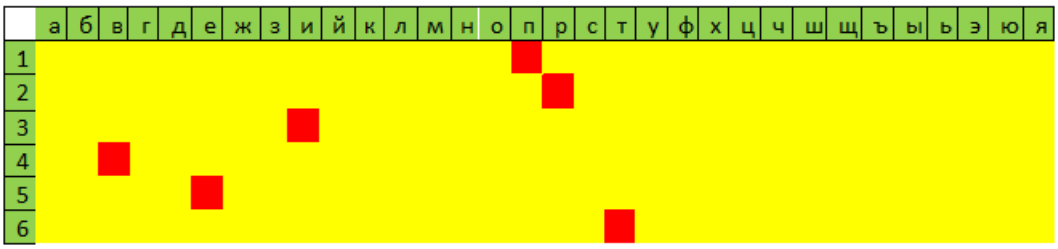


Рис. 2. Семантическое поле слова «привет»

Но возможен и другой вариант представления семантики текста – в виде двумерной матрицы или семантического поля. При этом колонки поля соответствуют отдельным семам, а строки – номеру семы в тексте. В качестве примера на рис. 2 в виде цветовой карты приведено семантическое поле для слова «привет». Здесь и далее изображения семантических полей приводятся только для диапазона русских букв.

В первоначальном состоянии все ячейки поля, соответствующие данной букве, равны единице, а остальные равны нулю.

Материалы и методы исследования

Одним из основных вопросов, связанных с созданием системы, является ее обучение. На начальных этапах «жизни» системы практически единственным методом обучения является обучение с учителем [3-5].

Алгоритмически этот процесс состоит из следующих этапов [6]:

1. На вход системы подается обучающая пара фраз (вопрос – ответ). Очевидным требованием к таким парам является практически полная завершенность диалога «вопрос – ответ».

2. Для введенной пары формируются соответствующие семантические поля.

3. Вопрос рассматривается системой как воздействие на нее внешней среды (в данном случае словесное). Ответ является реакцией системы на внешнее воздействие, которая нейтрализует это воздействие, т.е. семантические поля вопроса и ответа должны совпадать. Очевидно, что такое совпадение практически никогда не будет иметь места. Поэтому очередным этапом является корректировка указанных полей с целью достижения их равенства.

Для этого для каждой ячейки поля выполняется пересчет семантик по следующей цепочке формул:

сначала рассчитывается средняя семантика каждой пары ячеек:

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

где s_1 и s_2 – соответствующие семантики вопроса и ответа.

Далее определяется разница в семантике:

$$ds = |s_1 - s_2| \quad (3)$$

и затем производится корректировка семантик по формулам:

$$s_1 = \begin{cases} s_1 + ds/2, \ если\, s_1 < s_2 \\ s_1 - ds/2, \ если\, s_1 > s_2 \end{cases}$$
$$s_2 = \begin{cases} s_2 - ds/2, \ если\, s_1 < s_2 \\ s_2 + ds/2, \ если\, s_1 > s_2 \end{cases} \quad (4)$$

При этом отметим, что особенно для ранних стадий обучения составление обучающих пар может оказаться далеко не простой задачей.

Связано это с тем, что естественный язык обладает таким свойством, как вариативность, т.е. вопросы и ответы с одинаковым содержанием и одинаковые по смыслу можно задать с помощью различных фраз. В качестве примера в таблице приведены возможные варианты пар вопросов и ответов для выяснения имени человека.

Варианты вопросов и ответов при формировании семантики запроса об имени человека

Варианты запросов об имени	Варианты ответов
как тебя звать?	Катя
каково твое имя?	меня зовут Катя
как тебя называть?	мое имя Катя
как тебя можно называть?	вы можете называть меня Катя
как к тебе обращаться?	

Поэтому при обучении на вход системы подается не какая-то одна пара фраз, а поочередно все возможные комбинации пар. При этом получаются обобщенные семантические поля как вопросов, так и ответов, и их приходится постоянно корректировать по уравнению (2).

Изображение обобщенных семантических полей запросов и ответов об имени показано на рис. 3.

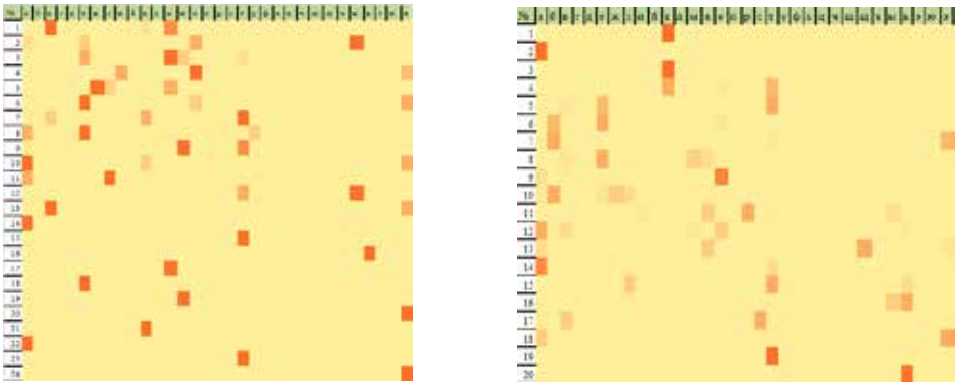
Вид получаемых полей зависит от порядка подачи фраз вопросов и ответов. Приведенные на рис. 3 поля получены при том порядке ввода, который указан в таблице. При этом часть ячеек полей принимают произвольные значения из диапазона от нуля до единицы. Необходимо также отметить, что при многократной подаче одних и тех же фраз все значения ячеек полей становятся близкими к единице и порядок подачи перестает иметь какое-то значение.

После получения обобщенной семантики вопросов и ответов с помощью уравнений (2) – (4) производится ее корректировка. Результат корректировки показан на рис. 4.

В полученных таким образом полях строки, соответствующие отдельным словам, и являются обобщенной семантикой каждого конкретного слова. Эта информация используется затем для формирования двух баз данных (БД) – БД слов-вопросов и БД слов-ответов.

Для записи слов-ответов привлекается информация о структуре предложений [7].

Согласно [7] любую фразу на естественном языке можно представить в виде некоторой структуры. Например, предложение «полученный спектр характеризует смысл обрабатываемого текста» (взято из данной публикации) можно представить в виде следующей структуры – рис. 5.При таком представлении во фразе, во-первых, выделяются группы подлежащего, сказуемого, дополнений и т.д., и, во-вторых, показываются отношения их подчиненности. Для лингвистов подобное представление – это просто удобный и наглядный способ показать взаимоотношения между словами в предложении.



а) *б)*

Рис. 3. Вид семантических полей для фраз-запросов (а) и фраз-ответов (б), представленных в таблице

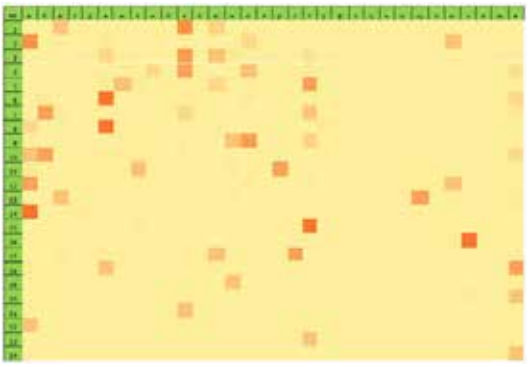


Рис. 4. Откорректированное семантическое поле запросов и ответов об имени человека

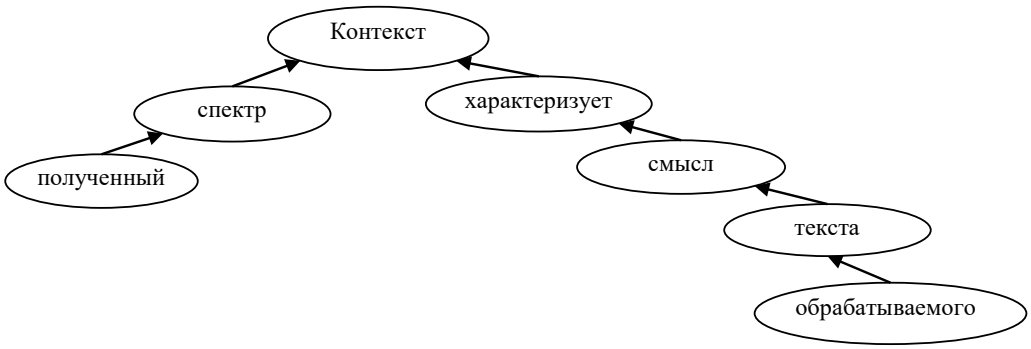


Рис. 5. Представление структуры фразы в виде дерева отношений

В данном же случае отношения между словами показывают порядок слияния семантических полей отдельных слов в общее семантическое поле фразы. При этом различный порядок слияния обеспечивает уникальность итоговой семантики фразы.

Поэтому в БД слов-вопросов записывается информация следующей структуры:

```
{
UpField: tField;
SlovoForma: string;                    (5)
SlovoField: tField;
}
```

где UpField – семантическое поле вышележащей словоформы;

SlovoForma – текущее слово;

SlovoField – семантическое поле текущего слова.

Отметим также, что запись результатов обучения в виде указанной структуры автоматически реализует и такое свойство естественного языка, как полисемия, т.е. зависимость смысла слова от лексического окружения.

В этом отношении показательна ситуация со словом «катя». При комбинировании различных вариантов вопросов и ответов это слово двадцать раз встречается в различных по структуре фразах, и оно находится на разных местах с различной собственной и предшествующей семантикой. Но в БД оно записывается только один раз в усредненном по семантике варианте, поскольку смысл всех двадцати комбинаций «вопрос – ответ» одинаковый. Если же данное слово встречается во фразах с другим смыслом, то оно записывается в БД дополнительно, но уже с другой семантикой.

Для записи слов-ответов формируется следующая структура:

```
{
QuestionField: tField;
SlovoForma: string;                    (6)
SlovoField: tField;
}
```

где QuestionField – семантическое поле запроса;

SlovoForma – слово;

SlovoField – семантическое поле данного слова.

Результаты исследования и их обсуждение

Сформированные таким образом БД используются затем при самостоятельной работе системы. Например, при реализа-

ции диалога работа происходит по следующей схеме:

1. В систему вводится вопрос.

2. Текст вопроса разбивается на слова.

3. Для каждого слова из БД считывается информация о его собственной и предшествующей семантике. Если таких словоформ несколько, то считываются все.

4. Для формирования семантики вопроса строится дерево распознавания.

4.1. Для этого берется первое слово и в тексте ищется слово, для которого предшествующая семантика текущего слова наиболее согласована с семантикой искомого слова.

4.2. Семантика таких слов обобщается.

4.3. Берется следующее, не использованное слово и повторяются операции 4.1–4.2.

4.4. Операции 4.1–4.3 повторяются до тех пор, пока не будут обработаны все слова текста вопроса.

5. Формируется ответ системы.

5.1. Для этого берется семантика запроса и в БД слов-ответов ищется слово, с наиболее близкой семантикой по полю QuestionField (6).

5.2. Из семантики запроса вычитается семантика найденного слова. Результат вычитания рассматривается как новый запрос.

5.3. Операции 5.1, 5.2 повторяются до тех пор, пока не будет полностью обнулена семантика запроса.

Из приведенной схемы следует, что как процесс распознавания смысла текста, так и процесс генерации ответа носят вероятностный характер, поскольку невозможно обеспечить точное равенство семантических полей слов предложения и слов, находящихся в БД.

Тем не менее эксперименты показывают, что система достаточно хорошо распознает смысл вопросов даже при относительно небольшом размере БД слов-вопросов (если, конечно, слова относятся к той предметной области, для которой система обучалась).

Намного сложнее обстоит ситуация с генерацией ответов.

Если речь идет о диалоге с технической командной системой, то достаточно будет провести обучение, в котором реализованы все возможные варианты ответов [8; 9]. Тогда генерация ответа будет сводиться к поиску в БД ответов записи, семантика которой наиболее согласована с семантикой вопроса.

Если же планируется создание системы, способной к ведению произвольных диалогов, то для получения осмысленных ответов необходимо создавать БД ответов по размерам, существенно большим, чем БД запросов.

Заключение

В публикации предложен новый подход к созданию интеллектуальных систем на основе понятия семантических полей. Введенное понятие позволяет учитывать не только наличие в тексте определенных символов, но и порядок их следования. При этом возможно построение семантических полей не только для отдельных слов, но и для текстов любых размеров.

Описана процедура начального обучения системы. На примерах показаны основные этапы обучения и особенности получаемых результатов. Одной из главных особенностей результатов является их соответствие данным лингвистики. В частности, это явление полисемии и зависимость смысла слов от лексического окружения.

Также показано, что данные, полученные в результате обучения, могут быть использованы для распознавания смысла текста и генерации ответов. Описанные при этом процедуры распознавания и генерации текстов могут быть использованы для создания простейших типов диалоговых систем.

Список литературы

1. Ванюлин А.Н., Алексеева Н.Р., Мочалова Т.А. Лингвистические основы алгоритмов компьютерной обработки текстов на основе систем с предопределенной семантикой // Современные наукоемкие технологии. 2020. № 3. С. 35–39.

2. 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.

3. Лимановская О.В., Алферьева Т.И. Основы машинного обучения: учебное пособие. Екатеринбург: изд-во Урал. ун-та, 2020. 88 с.

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

5. Вьюгин В.В. Математические основы теории машинного обучения и прогнозирования. М., 2013. 387 с.

6. Vanyulin A.N., Zvereva E.A., Lavina T.A., Mochalova T.A., Alekseeva N.R. Realization Algorithms of Major Types of Linguistic Processors Based on the Systems with Predefined Semantics. 2020 Global Smart Industry Conference (GloSIC), Chelyabinsk. 2020. P. 132-138. DOI: 10.1109/GloSIC50886.2020.9267857.

7. Касаткин Л.Л., Клобуков Е.В., Крысин Л.П., Лекант П.А. 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. Горковенко Д.К. Применение методов text mining для классификации информации, распространяемой в социальных сетях // Молодой ученый. 2016. №14 (118). С. 66–72. URL: https://moluch.ru/archive/118/32878/ (дата обращения: 10.03.2022).

9. 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: 20.01.2022).