



<http://dx.doi.org/10.35596/1729-7648-2020-18-4-44-52>

*Оригинальная статья*  
*Original paper*

УДК 004.774.2

## АВТОМАТИЧЕСКОЕ ПОСТРОЕНИЕ СЕМАНТИЧЕСКОЙ СЕТИ ДЛЯ ПОЛУЧЕНИЯ ОТВЕТОВ НА ВОПРОСЫ

ПОТАРАЕВ В.В., СЕРЕБРЯНАЯ Л.В.

*Белорусский государственный университет информатики и радиоэлектроники  
(г. Минск, Республика Беларусь)*

*Поступила в редакцию 4 марта 2020*

© Белорусский государственный университет информатики и радиоэлектроники, 2020

**Аннотация.** Выполнен анализ модели представления данных и знаний в виде семантической сети. Обоснован ее выбор для работы с текстовой информацией. Сформулирована задача автоматической генерации семантической сети на основе произвольного текста на русском языке. Приведены исходные данные, условия и ограничения, необходимые для алгоритма построения сети. В результате анализа части речи каждого слова и порядка слов в предложении установлены семантические отношения между словами. Создан словарь лексем, с помощью которого определяются части речи слов из предложений. Выбрано множество видов вопросов, используемых в семантической сети. Количество отношений в сети регулируется за счет возможности использовать только нужные типы связей при решении конкретной задачи. При этом отношения в семантической сети могут быть самых разных типов, что делает ее универсальной моделью представления данных и знаний. Разработан алгоритм, позволяющий получать ответы на поставленные вопросы. Рассмотрены предложения, для которых автоматически построена модель семантической сети. В предложенном алгоритме семантическая сеть интерпретируется как неориентированный граф, на котором для поиска ответа на вопрос применен алгоритм поиска в ширину. Разработанные алгоритмы реализованы в программном средстве, которое автоматически строит семантическую сеть для произвольного текста. Созданное программное средство позволяет задавать вопросы и получать на них ответы на основе информации, хранящейся в семантической сети. Эксперименты показали, что построенная семантическая сеть дает правильные ответы на поставленные ей вопросы. Сеть модифицируется путем добавления и удаления из нее информации. Есть возможность выбирать сложность структуры сети в зависимости от решаемой задачи. Предложенный подход к построению и работе с сетью позволяет использовать ее для текстов на разных языках, в информационных системах с естественно-языковым интерфейсом, для решения задач классификации и поиска информации.

**Ключевые слова:** семантическая сеть, вид вопроса, тип связи, автоматическая генерация, алгоритм поиска ответа.

**Конфликт интересов.** Авторы заявляют об отсутствии конфликта интересов.

**Для цитирования.** Потараев В.В., Серебряная Л.В. Автоматическое построение семантической сети для получения ответов на вопросы. Доклады БГУИР. 2020; 18(4): 44-52.

## AUTOMATIC GENERATION OF SEMANTIC NETWORK FOR QUESTION ANSWERING

VICTOR V. POTARAEV, LIYA V. SEREBRYANAYA

*Belarusian State University of Informatics and Radioelectronics (Minsk, Republic of Belarus)*

*Submitted 4 March 2020*

© Belarusian State University of Informatics and Radioelectronics, 2020

**Abstract.** Semantic network model for representing data and knowledge was analysed. Selection of this model for working with text information was justified. The objective of automatic semantic network generation based on an arbitrary Russian-language text was formulated. Initial data, conditions and constraints necessary for network generation algorithm are listed. As a result of the part-of-speech analysis for each word and word order in a sentence, semantic relations between words are determined. The Lexeme dictionary was created to determine the part of speech of words in sentences. A set of question types used in the semantic network was selected. The number of relations in the network is regulated due to the possibility to use only necessary relation types when resolving a specific task. With that, the relations in semantic network can have very different types, which makes it a universal model for representing data and knowledge. The algorithm was developed which allows one to get answers for the questions asked. The semantic network model was generated automatically for the sentences considered. In the proposed algorithm the semantic network is interpreted as unoriented graph on which breadth-first search algorithm is used to find an answer. The proposed algorithms were implemented in a software tool which automatically generates the semantic network for an arbitrary text. The created software tool allows asking questions and getting answers to them based on the information which is stored in the semantic network. The experiments have shown that the generated semantic network gives correct answers to the questions posed. The network is modified by adding and removing information in it. There is a possibility to choose complexity of network structure depending on a specific task being resolved. The proposed approach for building and working with the semantic network allows one to process texts in various languages, to use it in information systems with natural-language interface, and to resolve such tasks as text classification and text search.

**Keywords:** semantic network, question type, relation type, automatic generation, semantic analysis, question answering algorithm.

**Conflict of interests.** The authors declare no conflict of interests.

**For citation.** Potaraev V.V., Serebryanaya L.V. Automatic generation of semantic network for question answering. Doklady BGUIR. 2020; 18(4): 44-52.

### Введение

В современных информационных системах накоплено большое количество текстовой информации. В связи с этим актуальной является автоматизированная обработка текстов. Одним из инструментов, позволяющих обрабатывать текст с учетом его смысла, является семантическая сеть.

Семантическая сеть – это ориентированный граф, который отражает понятия и отношения между ними [1]. Вершины графа содержат понятия, а связи графа – это отношения между понятиями. Типы связей в графе семантической сети выбираются создателем сети в зависимости от конкретных целей [2]. Также отношения в сетях могут быть разных типов: функциональные, количественные, пространственные, временные, логические и др. Широко применяются иерархические семантические сети, имеющие древовидную структуру. Семантическая сеть является моделью, которая чаще всего используется при создании баз знаний [3]. Данная модель может хранить текстовую информацию в узлах сети без привлечения дополнительного хранилища, что упрощает решение задач, связанных с обработкой текстов. Кроме того, такая модель соответствует научным представлениям об организации семантической памяти человека [4].

Семантические сети часто используются в системах искусственного интеллекта и обработки естественного языка, в вопросно-ответных и предметно-ориентированных системах. Примерами таких систем являются *WordNet* и *PyTez*, созданные для обработки англоязычной и русскоязычной информации соответственно [5].

*WordNet* – это лексическая база данных, содержащая словарь и набор семантических сетей. Она применяется для информационного поиска, в работе вопросно-ответных систем и др. Преимущества модели в виде сети: разнообразие видов отношений; поддержка многозначности слов. Недостатки: огромный размер сетей; трудоемкое создание экспертом вручную; слова разных частей речи не связаны между собой; ориентированность на английский язык.

*PyTez* – это словарь русского языка, представляющий собой иерархическую сеть понятий, для которых указаны текстовые выражения. Сеть применяется для информационного поиска, автоматического расширения запроса, автоматической рубрикации и др. Преимущества используемой модели: поддержка многословных понятий; понятие может включать в себя слова разных частей речи; возможность модификации. Недостатки: огромный размер сети; трудоемкое создание экспертом вручную; ориентированность на русский язык.

В современных языках существует большое количество неологизмов, и оно постоянно растет. Некоторые слова получают новый смысл. Из-за этого производится все больше попыток автоматизации создания семантических сетей. При создании сетей используется структурированная и неструктурированная информация [6]. Могут быть использованы словари, тексты на разных языках и машинный перевод [7]. Существует множество автоматически сгенерированных семантических сетей, используемых для различных языков, например *BabelNet* [8].

*BabelNet* – это семантическая сеть, которая сочетает в себе связи из *WordNet* с семантическими отношениями, построенными на основе статей Википедии. Применяется для вычисления семантической близости между понятиями, для определения используемого значения слова и др. Преимущества этой модели сети: автоматическое построение; разнообразие видов отношений; поддержка многозначности слов; поддержка различных языков. Недостатки: огромный размер; в основе лежат связи из *WordNet*, полученные вручную.

Анализ популярных семантических сетей показал, что актуальной остается задача автоматического построения сети с последующим ее использованием.

Целью данной работы является создание алгоритма автоматической генерации семантической сети, которая может быть применена для получения ответов на вопросы на русском языке. Размер и структура сети зависят от решаемой задачи. Созданная сеть может быть использована в системах с естественно-языковым интерфейсом, которые интенсивно развиваются на протяжении нескольких последних десятилетий [9].

### Алгоритм автоматического построения семантической сети

Рассмотрим задачу автоматической генерации семантической сети, используемой для нахождения ответов на вопросы, выраженные на естественном языке.

При решении поставленной задачи предлагается использовать следующие исходные данные:

1. Текст, для которого необходимо построить сеть.
2. Словари частей речи русского языка.
3. Словарь синонимов.
4. Список поддерживаемых видов вопросов.

В русском языке существуют различные виды вопросов – закрытые, открытые, переломные вопросы и др. Предполагается, что построенная семантическая сеть сможет отвечать на открытые вопросы. Примеры таких вопросов: «где?», «как?», «кто?». Типы связей, используемые в семантической сети, зависят от видов вопросов, которые необходимо поддерживать. Кроме того, для построения предложения в ответ на запрос пользователя в семантической сети необходимо наличие связи «подлежащее-сказуемое».

Для корректной работы алгоритма требуется выполнение следующих условий:

1. Типы связей, имеющиеся в семантической сети, позволяют найти ответ на вид вопроса, который поставлен.

2. В тексте отсутствуют многозначные слова.

3. Семантические отношения между словами текста, знание о которых необходимо для ответа на вопрос, можно определить автоматически. Это означает, что в случае применения простейшего алгоритма определения членов предложения в каждом предложении текста подлежащим должно являться существительное в начальной форме, а сказуемым – глагол.

При нахождении основы слова может быть применен один из алгоритмов стемминга, например, алгоритм Портера.

При построении семантической сети будем искать слова, относящиеся к конкретным частям речи. Для определения того, к какой части речи принадлежит слово, можно сравнить основу данного слова с основами слов, чья принадлежность уже установлена. Существуют более универсальные и/или точные способы определения частей речи, но возможностей выбранного подхода достаточно для обработки несложного текста.

Алгоритм построения семантической сети может быть представлен следующим образом:

1. Определить типы связей, необходимые для решения поставленной задачи (в данном случае – для ответа на поддерживаемые виды вопросов).

2. Перейти к очередному предложению текста, начиная с первого.

3. Добавить в сеть связи «подлежащее – сказуемое». В качестве подлежащего можно использовать существительное в начальной форме. В качестве сказуемого – глагол.

4. Добавить в сеть типы связей, определенные на шаге 1.

5. Добавить в сеть связи «слово – основа слова».

6. Добавить в сеть связи типа «синоним», соединяющие основы слов.

7. Повторять шаги 3–6 для каждого предложения до конца текста.

В случае необходимости добавления информации в уже существующую сеть требуется повторять шаги 3–6 для новых предложений.

Для удаления неактуальной или устаревшей информации нужно найти связи, соответствующие предложениям текста с этой информацией, и удалить их из сети. Если при этом появляются узлы, не имеющие связей, то их рекомендуется тоже удалить.

### Пример автоматического построения семантической сети

Даны три предложения: «На стадионе проводится матч», «Из-за матча в городе опустели улицы», «Скоро будет многолюдно». Покажем, что можно построить модель семантической сети, позволяющую отвечать на виды вопросов, перечисленные в табл. 1.

Таблица 1. Виды вопросов  
Table 1. Question types

Вид вопроса Question type	Что выступает в качестве ответа What is an answer	Признак, который можно использовать для поиска ответа Word characteristic which can be used to find the answer
Кто/Что?	Объект, выполняющий действие	Существительное в начальной форме
Где?	Указание места	Пространственный предлог + существительное не в начальной форме
Как?	Указание, каким образом выполнено действие	Наречие образа действия
Из-за чего?	Указание, по какой причине выполняется действие	Предлог «из-за» + существительное не в начальной форме

Пусть списки (словари) заранее известных слов выглядят, как показано в табл. 2.

Рассмотрим определение части речи на примере. Для слова «улица» в словаре существительных при помощи алгоритма стемминга получаем основу «улиц». Для слова

«улице» из предложения «Дождь идет на улице» после удаления окончания «е» получим также основу «улиц». Следовательно, «улице» – это существительное. Аналогичным образом выполняется анализ того, какие суффиксы и окончания могут быть у глаголов.

Для слова «солнечно» в предложении «Завтра будет солнечно» после удаления окончания «о» останется основа «солнечн», которая не совпадает с основой ни одного существительного. Значит, это не существительное. Более того, слово «солнечно» находится в словаре наречий образа действия.

Слова «быть» и «будет» – разные формы одного слова, но при помощи предложенного подхода это определить нельзя. Поэтому слово «будет» добавлено в список глаголов.

**Таблица 2.** Списки заранее известных слов  
**Table 2.** Lists of pre-known words

Вид списка List type	Содержимое списка List contents
Существительные	Дождь, улица, лужа, асфальт
Глаголы	Проводить, опустеть, быть, будет
Пространственные предлоги	В, на (для решения данного примера достаточно знания об одном предлоге)
Причинные предлоги	Из-за (для построения более полной модели семантической сети в данном случае необходимо знание об этом предлоге)
Наречия образа действия	Многолюдно (для решения данного примера достаточно знания об одном наречии)
Словарь синонимов	Матч – игра, улица – дорога

Рассмотрим работу алгоритма автоматического построения семантической сети на примере трех выбранных предложений.

1. Пусть сеть будет использоваться для ответа на вопросы: «Где?», «Как?», «Из-за чего?». Этим вопросам соответствуют типы связей «место», «образ действия», «причина действия».

2. Первое предложение – «На стадионе проводится матч».

3. Подлежащим является существительное в начальной форме «матч», сказуемым – глагол «проводится».

Если в предложении обнаружено несколько подлежащих или сказуемых, то в сети для данного предложения будет несколько связей «подлежащее – сказуемое».

4. Добавим в сеть типы связей, определённые на шаге 1.

а. Место (соответствует вопросу «Где?»).

На данном этапе (согласно табл. 1) надо найти в предложении пару «Пространственный предлог + существительное не в начальной форме». Добавим «на стадионе» в качестве обстоятельства места.

б. Образ действия (соответствует вопросу «Как?»). В этом предложении нет наречий образа действия. Сеть не изменяется.

с. Причина действия (соответствует вопросу «Из-за чего?»). На данном этапе надо найти в предложении пару «Причинный предлог + существительное не в начальной форме». Причинные предлоги в предложении не найдены. Сеть не изменяется.

5. Для существительных и глаголов из предложения получаем основы слов: «на стадионе» – «стадион», «проводится» – «провод», «матч» – «матч». Слово «матч» совпадает со своей основой. Два узла с одинаковым текстом являются избыточными для поиска ответа на вопрос. Поэтому новый узел и связь не добавляются.

6. Используя словарь синонимов, определяем, что слова «матч» и «игра» – синонимы. Получим сеть, представленную на рис. 1.



1. Подлежащее вопроса – существительное в начальной форме «игры». Основа подлежащего – «игр». Сказуемое – «проводятся», основа которого «провод», есть в сети.

Во время поиска фрагмента сети, соединяющего слова вопроса, неважно, в каком направлении указана связь. Направление используется для определения роли связанных слов и не влияет на тот факт, что слова связаны по смыслу. Поэтому семантическую сеть можно рассматривать как неориентированный граф. Применим для ответа на вопрос алгоритм поиска в ширину. В результате будет найден путь «провод» – «проводится» – «матч» – «игр».

2. В найденном фрагменте сети дублируются слова: синонимы и разные формы одного слова. Для формирования ответа, понятного пользователю, он строится из слов исходного текста. Последовательно анализируем связи между узлами найденного фрагмента сети. В найденном пути первый узел «провод» связан со вторым узлом «проводится» связью «слово – основа слова», поэтому он отбрасывается. У слова «проводится» две связи (с предыдущим и последующим узлами найденного фрагмента), одна из которых – «подлежащее – сказуемое». Слово добавляется в ответ.

У слова «матч» также две связи, одна из которых – «подлежащее – сказуемое», и это слово также добавляется в ответ. Последнее слово «игр» связано с предыдущим словом только связью «синоним» и поэтому отбрасывается.

В результате получим два слова ответа: проводится матч.

3. Вопрос «Где проводятся игры?» подразумевает определение места. Место, связанное со словами найденного фрагмента сети «на стадионе», добавляется в ответ. В результате получен ответ: «Проводится матч на стадионе».

Если для найденного фрагмента сети невозможно определить место, то это означает, что ответ на вопрос не удастся найти.

Аналогичным образом можно показать, что описанный алгоритм позволяет найти ответы на другие вопросы, например:

- 1) «Где опустели улицы?» – «Опустели улицы в городе»;
- 2) «Как будет потом?» – «Будет многолюдно»;
- 3) «Из-за чего улицы опустели?» – «Улицы опустели из-за матча».

Нахождение пути на графе может производиться с применением поиска в ширину, поиска в глубину или другого алгоритма.

### Результаты исследований

Разработанные алгоритмы были реализованы в программном средстве, которое автоматически строит семантическую сеть для произвольного текста. Созданное программное средство позволяет задавать вопросы и получать на них ответы, основанные на информации, хранящейся в семантической сети.

Проверка сети показала, что для рассмотренных примеров, а также для других несложных текстов, она дает верные ответы. В случае обработки текста с новыми словами необходимо лишь расширить списки известных слов, например, на основе словарей русского языка.

Предложенный подход с небольшими изменениями может быть реализован не только для русского языка, но и для других языков. Например, для английского языка при необходимости определения подлежащего стоит учитывать не форму слова, а порядок слов в предложении, однако в целом подход не меняется.

Для поиска ответа на вопрос могут быть использованы и другие модели представления информации, например, искусственные нейронные сети. Однако преимуществом модели семантической сети является ее ориентированность на обработку естественного языка. Она не требует преобразования текстовой информации в другой вид, а также удобна для восприятия человеком. Добавление предложений приводит к расширению сети, и количество сохраняемых предложений не ограничено. Проведенные эксперименты показали, что нахождение ответа на вопрос с применением семантической сети – это хорошо формализуемая и автоматизируемая процедура, а ее результат значительно зависит от способа генерации сети и корректности определения семантических отношений слов.

## Заключение

В работе предложены алгоритмы автоматической генерации семантической сети на основе произвольного текста, а также поиска ответов на заданные вопросы с помощью полученной сети. Алгоритмы предназначены для работы с несложными текстами, где каждый член предложения должен быть представлен предусмотренными частями речи.

Преимуществами предложенного способа генерации модели являются автоматическое построение сети для произвольного текста, а также возможность выбора сложности структуры сети в зависимости от решаемой задачи. Планируется усовершенствование алгоритма за счет увеличения числа видов отношений между понятиями, добавления поддержки многозначных слов, а также обработки текстов на других языках.

Модель семантической сети, предложенная в данной статье, может быть положена в основу работы информационных систем с естественно-языковым интерфейсом. Кроме того, подобные модели нашли свое применение для решения задач классификации и поиска текстовой информации [10].

## Список литературы

1. Гаврилова Т.А. *Базы знаний интеллектуальных систем*. СПб.: Питер; 2000.
2. Рахимова Д.Р. Построение семантических отношений в машинном переводе. *Вестник КазНУ им. аль-Фараби. Серия: Математика, механика и информатика*. 2014;80(1):90-101.
3. Овчиева Ю.А. Семантическая сеть – перспективная платформа для системы управления знаниями. *Вестник университета*. 2015;3:14-16.
4. Осипов Г.С. *Методы искусственного интеллекта*. Москва: Физматлит; 2011.
5. Лукашевич Н.В. *Тезаурусы в задачах информационного поиска*. Москва: Издательство МГУ; 2011.
6. Усталов Д.А., Созыкин А.В. Комплекс программ автоматического построения семантической сети слов. *Вестник ЮУрГУ. Серия: Вычислительная математика и информатика*. 2017;6(2):69-83.
7. Wong W. Ontology Learning from Text: A Look Back and into the Future. *ACM Computing Surveys*. 2012;44(4):20:1-20:36.
8. Navigli R., Ponzetto S.P. BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. *Artificial Intelligence*. 2012;193:217-250.
9. Bouziane A., Bouchiha D., Doumi N., Malki M. Question Answering Systems: Survey and Trends. *Procedia Computer Science*. 2015;73:366-375.
10. Серебряная Л.В., Потараев В.В. Методы классификации текстовой информации на основе искусственной нейронной и семантической сетей. *Информатика*. 2016;52(4):95-103.

## References

1. Gavrilova T.A. [*Knowledge bases of intellectual systems*]. SPb.: Piter; 2000. (In Russ.)
2. Rahimova D.R. [Creation of the semantic relations in machine translation]. *Vestnik KazNU im. al'-Farabi. Seriya: Matematika, mekhanika i informatika=Journal of Mathematics, Mechanics and Computer Science*. 2014;80(1):90-101. (In Russ.)
3. Ovchieva J.A. [Semantic web – a promising platform for knowledge management system]. *Vestnik universiteta= Vestnik universiteta*. 2015;3:14-16. (In Russ.)
4. Osipov G.S. [*Methods of artificial intelligence*]. Moscow: Fizmatlit; 2011. (In Russ.)
5. Lukashevich N.V. [*Thesauruses in information search tasks*]. Moscow: Publishing House of MSU; 2011. (In Russ.)
6. Ustalov D.A., Sozykin A.V. [A Software System for Automatic Construction of a Semantic Word Network]. *Vestnik YuUrGU. Seriya: Vychislitel'naya matematika i informatika = Bulletin of the South Ural State University. Series: Computational Mathematics and Software Engineering*. 2017;6(2):69-83. (In Russ.)
7. Wong W. Ontology Learning from Text: A Look Back and into the Future. *ACM Computing Surveys*. 2012;44(4):20:1-20:36.
8. Navigli R., Ponzetto S.P. BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. *Artificial Intelligence*. 2012;193:217-250.
9. Bouziane A., Bouchiha D., Doumi N., Malki M. Question Answering Systems: Survey and Trends. *Procedia Computer Science*. 2015;73:366-375.
10. Serebryanaya L.V., Potaraev V.V. [Methods of textual information classification based on artificial neural and semantic networks]. *Informatika=Informatics*. 2016;52(4):95-103. (In Russ.)

### Вклад авторов

Потараев В.В. разработал алгоритм построения семантической сети и алгоритм ответа на вопрос, выполнил экспериментальную проверку разработанных алгоритмов.

Серебряная Л.В. сформулировала задачи, которые необходимо было решить в ходе исследований, а также выполнила анализ и интерпретацию полученных результатов.

### Authors' contribution

Potaraev V.V. developed the algorithm of semantic network generation, the algorithm of question answering, and performed their experimental validation.

Serebryanaya L.V. identified the tasks to be solved during the research, and also participated in the interpretation of their results.

### Сведения об авторах

Потараев В.В., м.т.н., аспирант кафедры программного обеспечения информационных технологий Белорусского государственного университета информатики и радиоэлектроники.

Серебряная Л.В., к.т.н., доцент, доцент кафедры программного обеспечения информационных технологий Белорусского государственного университета информатики и радиоэлектроники.

### Information about the authors

Potaraev V.V., M.Sci., PG student of Information Technologies Software Department of Belarusian State University of Informatics and Radioelectronics.

Serebryanaya L.V., PhD, Associate Professor, Associate Professor of Information Technologies Software Department of Belarusian State University of Informatics and Radioelectronics.

### Адрес для корреспонденции

220013, Республика Беларусь,  
г. Минск, ул. П. Бровки, 6,  
Белорусский государственный университет  
информатики и радиоэлектроники  
тел. +375-17-293-84-93;  
e-mail: L\_silver@mail.ru  
Серебряная Лия Валентиновна

### Address for correspondence

220013, Republic of Belarus,  
Minsk, P. Brovka str., 6,  
Belarusian State University  
of Informatics and Radioelectronics  
tel. +375-17-293-84-93;  
e-mail: L\_silver@mail.ru  
Serebryanaya Liya Valentinovna