**Инженерный вестник Дона**, № 4 (2015)
ivdon.ru/ru/magazine/archive/n4y2015/3304

# О возможном подходе к представлению денотатной структуры предметной области в системах автоматического реферирования

*Е.В. Долгова[1], Д.С. Курушин[1], Н.М. Нестерова[1],*
*О.В. Соболева[1], А.Н. Паньков[2], М.И. Хакимова[2]*

[1]*Пермский национальный исследовательский политехнический университет, Пермь*
[2]*Пермский государственный национальный исследовательский университет, Пермь*

**Аннотация:** В статье рассматривается проблема представления структуры предметной области, которая могла бы использоваться в системах автоматического реферирования. В качестве решения данной проблемы авторы предлагают использовать методику денотатного анализа текста, разработанного А.И. Новиковым и его школой (Институт языкознания РАН). Данная методика легла в основу предлагаемой авторами инфологической модели, которая делает возможным формализованное представление структуры определенной предметной области («понятное» машине) и в конечном итоге создание системы автоматического реферирования.
Работа выполняется при поддержке РФФИ, проект №14-07-00671.
**Ключевые слова:** денотат, реферирование, понимание, инфологическая модель, нейронная сеть, понимание текста, смысловое свертывание.

## Проблематика и существующие решения

В настоящее время объем данных, используемых в научно-технической, социально-экономической, учебной, культурной и иных сферах вырос настолько, что породил проблему обесценивания представленной в них информации. Это обусловлено объективными трудностями поиска среди множества разнородных текстов. В итоге увеличение их количества не улучшает доступ к необходимой информации, а нередко ухудшает его, из-за невозможности провести структуризацию больших данных силами людей-специалистов. В такой ситуации актуальными становятся технологии и методы реферирования текстовой информации при помощи программных систем. Под реферированием в данном случае будем понимать создание краткого изложения текста, то есть, реферата. В работе Новикова А.И и Нестеровой Н.М. [1] реферирование определяется как один из видов аналитико-синтетической обработки первичных документов, результатом которой является создание вторичных документов – рефератов, а рефератом

© Электронный научный журнал «Инженерный вестник Дона», 2007–2015

является кратко изложенный первичный документ. Согласно определению В.И. Соловьева реферат должен обладать определенными свойствами – малым объемом, полнотой и постоянством структуры.

Практически все современные методы автореферирования основаны на квазиреферировании [2,3], при котором в тексте выделяются наиболее информативные фразы, а из них, в свою очередь, формируется квазиреферат. Следует также отметить, что некоторые общие и частные задачи, связанные с обработкой текстов или созданием вторичных документов, в настоящее время автоматизированы. Например, существует система TextAnalyst 2.0, которая позволяет анализировать содержание текста и формировать семантическую сеть с гиперссылками [4]. Однако, созданная этой системой семантическая сеть не используется ею для реферирования.

Таким образом, до сих пор не имеют полного решения те задачи формирования вторичных текстов, которые связаны с семантическим анализом и синтезом. В подобных случаях речь идет о понимании текста, который представляет собой двухсторонний знак, одной стороной этого знака является его внешняя (языковая) форма, непосредственно нами воспринимаемая, второй – внутренняя, (содержательная) , которая «скрыта» в тексте и которую нужно «найти». Если внешняя форма — это линейная последовательность языковых знаков, соединенных по особым формальным правилам, то внутренняя форма — это модель фрагмента некой предметной ситуации, образованной совокупностью предметов и их отношений. Она носит не линейный, а иерархический характер. Это означает, что процесс понимания можно представить как переход от внешней формы текста к внутренней. Основная трудность, которая возникает при попытке формализовать этот процесс, вызвана отсутствием изоморфизма между этими формами текста, между теми кодами, которыми они представлены: в первом случае — это естественный язык, во втором — универсальный

© Электронный научный журнал «Инженерный вестник Дона», 2007–2015

предметный код [5] , отражающий денотатный (денотат — от лат. denotatum — обозначаемое / обозначаемый объект) уровень знания о предметной действительности. Это означает, что содержанием текста является его денотатная структура, которая, являясь внутренним мыслительным образованием, не имеет своих собственных средств выражения, кроме средств естественного языка [6, с.129]. Таким образом, денотатная структура — это тоже форма текста, в которой отражено содержание. Однако для адекватного отражения содержания данная форма должна соответствовать определенным требованиям, а именно, требованиям целостности, пространственности, изобразительности, а также в ней должны быть обеспечены однозначность, эксплицитность, иерархичность элементов содержания. В качестве способа отражения денотатной структуры, удовлетворяющей названным требованиям, А.И. Новиковым была разработана методика построения денотатного графа, «где вершинам соответствуют имена денотатов, полученные в результате содержательного анализа текста и применения необходимых знаний о данном фрагменте действительности, а ребрам — предметные отношения между этими денотатами» [6, с.131]

Представление содержания в виде графа накладывает определенные ограничения на выбор языковых единиц для обозначения имен денотатов из отношений. Так, для имен денотатов используются только номинативные элементы языка, при этом выбираются те, которые являются инвариантными для обозначения объектов. Для обозначения межпредметных отношений используются глагольные конструкции, количество которых невелико.

Отметим, что вышеописанная задача является задачей искусственного интеллекта, поскольку одновременно связана с моделированием понимания, применением программной системы, учетом факторов неопределенности, отсутствием явных критериев оптимизации и т.д. Известно, что о

© Электронный научный журнал «Инженерный вестник Дона», 2007–2015

прохождении теста Тьюринга современными системами пока что говорить не приходится. Таким образом задача, которую А.И. Новиков считает неразрешимой, связана с другой неразрешенной в настоящее время задачей – с общей проблемой эквивалентности искусственной системы и естественного интеллекта специалиста.

Однако невозможность решения общей неограниченной задачи не означает невозможности решения частных задач с ограничениями (такая задача решается, например, в [7], или в [8]). Очевидно, переход от лексико-грамматической формы текста к его представленной в виде графа денотатной структуре может объектом автоматизации на основе моделей и методов, наработанных в сфере ИИ.

## Инфологическое моделирование

Представление смыслового преобразования как переход «Первичный текст — Денотатная структура — Вторичный текст» дает основание для условного выделения в этом процессе двух дискретных этапов. Первый этап заключается в смысловом свертывании текста. Второй этап представляет собой развертывание денотатной структуры в соответствующий вторичный текст, который должен быть семантически адекватным первичному.

Второй этап смыслового преобразования текста тривиален, получить вторичный текст на основе денотатной структуры не сложно. Первый этап – более сложный и включает в себя несколько задач. Во-первых, необходимо разработать инфологическую модель для хранения денотатной структуры предметной области, во-вторых требуется спроектировать искусственную нейронную сеть, которая будет выделять денотаты из исходного текста. Наконец, в третьих, необходимо выбрать способ визуализации денотатного графа для промежуточной проверки качества работы нейронной сети

© Электронный научный журнал «Инженерный вестник Дона», 2007–2015

человеком-специалистом, такая сеть по структуре и методам обучения и приблизительно эквивалентна сетям, представленным в работе [9].

На рисунке 1 представлен полученный в работе результат — диаграмма «сущность-связь» инфологической модели. База данных в этом случае используется как оболочка для сознания базы знаний интеллектуальной системы, что является широко распространенным подходом.

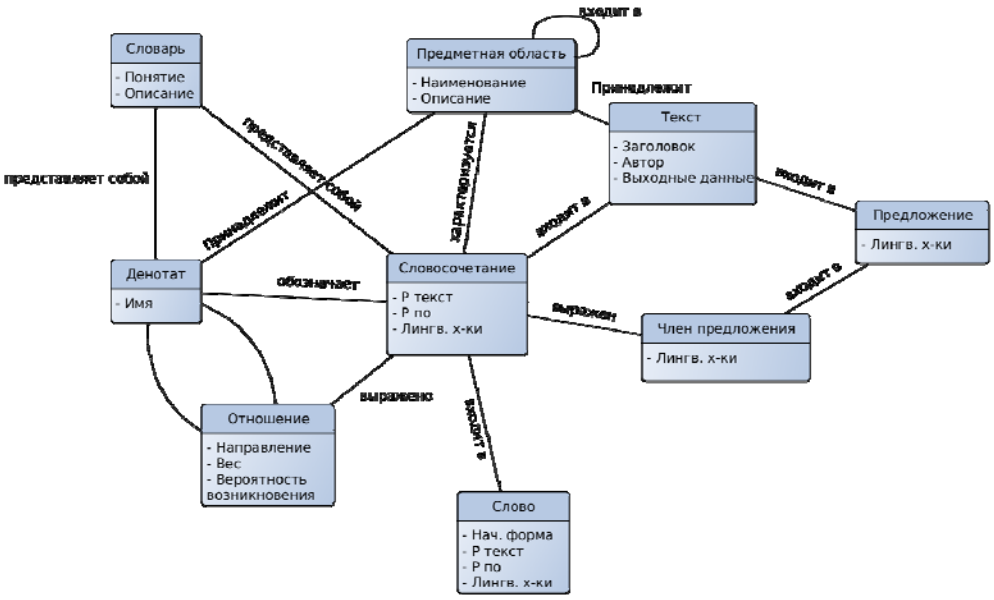


Рис. 1. – Диаграмма «сущность-связь» (фрагмент)

На рисунке 2 представлен пример денотатного графа, генерируемый системой. Прямоугольниками обозначены денотаты, овалами — отношения.

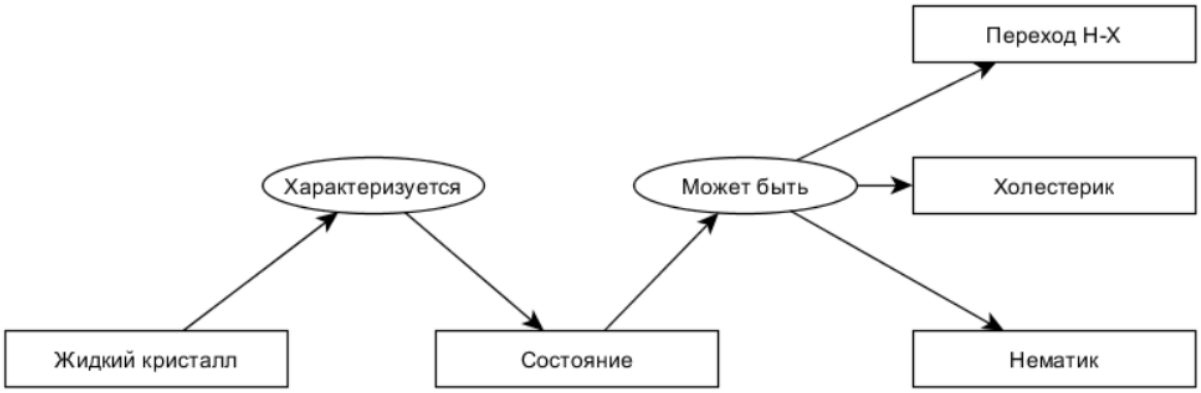


Рис. 2. – Пример денотатного графа

© Электронный научный журнал «Инженерный вестник Дона», 2007–2015

Подробное рассмотрение нейронной сети-распознавателя выходит за рамки данной работы, однако, общие ее характеристики сводятся к следующим:

1) входным сигналом сети является текст (упорядоченный набор слов), единовременно на вход сети поступает фрагмент исходного текста, размер которого ограничен размерностью сети, при этом предложения на части не разбиваются;

2) выходным сигналом сети является кодированное представление искомого графа, который, в свою очередь, является множеством кортежей <объект1, объект 2, связь>;

3) обучение нейросети может осуществляться как с выборкой, так и без нее, второй метод гораздо менее трудоемкий, но чреват получением очень «нестандартного» с человеческой точки зрения аналитического результата.

В отличие от сходной задачи, решенной в [10], для визуализации графов используется Graphviz — пакет утилит для автоматической визуализации графов, заданных в виде описания языка DOT [11]. К достоинствам этого инструментария следует отнести поддержку многих операционных систем, возможность вывода графа во разные форматы (PNG, SVG, PDF), поддержку разнообразных моделей описания графа, простоту графического представления, хорошую совместимость со скриптовыми языками программирования.

## Выводы

Таким образом, результатом представленной работы является инфологическая модель, а также методика ее применения. Они, в свою очередь, являются базой, на основе которой возможна разработка системы автоматического реферирования текста с учетом его смысла.

© Электронный научный журнал «Инженерный вестник Дона», 2007–2015

Такая система не создаст идеального реферата для любого текста автоматически, но за счет привязки к определенной предметной области, интерактивного взаимодействия с человеком-специалистом, а также периодической перенастройки сети может дать вполне приемлемый результат. В итоге она способно существенно сократить трудозатраты на реферирование научно-технических текстов.

## Литература

1. Новиков А.И., Нестерова Н.М. Реферативный перевод научно-технических текстов. М.: Академия наук СССР, Институт Языкознания, 1991. 147 с.

2. Киселёв Ю.А. Перспективы использования жанровой классификации Веб документов в поисковых системах // Инженерный вестник Дона. 2012. №4. URL: ivdon.ru/ru/magazine/archive/n4p2y2012/1425.

3. Автореферат. URL: referat.keywordrush.com (дата обращения 02.10.2015).

4. TextAnalyst 2.0 – персональная система автоматического анализа текста. URL: analyst.ru/index.php?lang=eng&dir=content/products/&id=ta (дата обращения 02.10.2015).

5. Жинкин Н.И. Речь как проводник информации. М.: Наука, 1982. 156 с.

6. Новиков А.И. Семантика текста и ее формализация. М.: Наука, 1983. 214 с.

7. Файзрахманов Р.А., Файзрахманов Р.Р., Долгова Е.В. Моделирование представления информации в задачах автоматической обработки веб-страниц и извлечения веб-информации // Вестник Ижевского государственного технического университета. 2011. № 2. С. 176-178

© Электронный научный журнал «Инженерный вестник Дона», 2007–2015

8. Och F.J., Tillmann C., Ney H. Improved Alignment Models for Statistical Machine Translation. URL: ai.mit.edu/courses/6.891-nlp/ASSIGNMENT1/t1.4.pdf (accessed 02/10/2015).

9. Долгова Е.В., Файзрахманов Р.А. Выбор модели технической системы на основе технологии распознавания // Приборы и системы. 2005. № 9. С. 68-70.

10. Носко В.И. Система автоматизированного построения графа социальной сети // Инженерный вестник Дона. 2012. №4. URL: ivdon.ru/ru/magazine/archive/n4p2y2012/1428.

11. Graphviz - Graph Visualization Software. URL: graphviz.org (accessed 02/10/2015).

**References**

1. Novikov A.I., Nesterova N.M. Referativnyy perevod nauchno-tekhnicheskikh tekstov [Patent translation of scientific and technical texts]. Moscow: Akademiya nauk SSSR, Institut Yazykoznaniya, 1991. 147 p.

2. Kiselev Yu.A. Inženernyj vestnik Dona (Rus), 2012. №4. URL: ivdon.ru/ru/magazine/archive/n4p2y2012/1425.

3. Avtoreferat [Author's abstract]. URL: referat.keywordrush.com (accessed 02/10/2015).

4. TextAnalyst 2.0 – personal'naya sistema avtomaticheskogo analiza teksta [TextAnalyst 2.0 - personal system of automatic text analysis]. URL: analyst.ru/index.php?lang=eng&dir=content/products/&id=ta (accessed 02/10/2015).

5. Zhinkin N.I. Rech' kak provodnik informatsii [Speech as a conduit of information]. Moscow: Nauka, 1982. 156 p.

6. Novikov A.I. Semantika teksta i ee formalizatsiya [The semantics of the text and its formalization]. Moscow: Nauka, 1983. 214 p.

© Электронный научный журнал «Инженерный вестник Дона», 2007–2015

7. Fayzrakhmanov R.A., Fayzrakhmanov R.R., Dolgova E.V. Vestnik Izhevskogo gosudarstvennogo tekhnicheskogo universiteta. 2011. № 2. pp. 176-178

8. Och F.J., Tillmann C., Ney H. Improved Alignment Models for Statistical Machine Translation. URL: ai.mit.edu/courses/6.891-nlp/ASSIGNMENT1/t1.4.pdf (accessed 02/10/2015).

9. Dolgova E.V., Fayzrakhmanov R.A. Pribory i sistemy. 2005. № 9. pp. 68-70.

10. Nosko V.I. Inženernyj vestnik Dona (Rus), 2012. №4. URL: ivdon.ru/ru/magazine/archive/n4p2y2012/1428.

11. Graphviz - Graph Visualization Software. URL: graphviz.org (accessed 02/10/2015).

© Электронный научный журнал «Инженерный вестник Дона», 2007–2015