Инженерный вестник Дона, № 4 (2015)
ivdon.ru/ru/magazine/archive/n4y2015/3311

Экспериментальные исследования денотативной модели понимания в приложениях автоматического реферирования текста

Н.А. Герте, Д.С. Курушин, Н.М. Нестерова, О.В. Соболева

Пермский национальный исследовательский политехнический университет, Пермь

Аннотация: В статье рассматривается экспериментальное исследование компьютерного представления структуры предметной области, которая может быть использована в системе автоматического реферирования. В качестве теоретической основы исследования была выбрана психолингвистическая теория А.И. Новикова и разработанная им методика денотативного анализа текста, позволяющая эксплицировать в виде графа структуру как отдельного текста, так и определенной предметной области. Использование данной методики позволило авторам создать вычислительную модель для автоматического построения графов, отражающих содержание вводимых в машину текстов.
Работа выполняется при поддержке РФФИ, проект №14-07-00671.
Ключевые слова: денотат, вычислительный эксперимент, реферирование, понимание, инфологическая модель, понимание текста, смысловое свертывание.

Постановка проблемы и предлагаемое решение

Создание системы автоматизированного реферирования не является новой задачей [1, 2], но до сих пор она остается нерешенной. «Неразрешимость» этой задачи связана с тем, что для ее решения требуется найти способ формализации не внешней (языковой) формы текста, а внутренней (содержательной) [3, 4]. Это, в свою очередь, требует создания модели понимания, применимой в человеко-машинной коммуникации.

Представляется, что в качестве такой модели может быть использована модель содержания текста, представленная иконически в виде денотатного графа, отражающего иерархическую систему денотатов и их отношений, что соответствует модели фрагмента реальной предметной ситуации. Методика построения такого графа, в котором «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» была разработана в отечественной лингвистике А.И. Новиковым [5, с.131].

В рамках выполнения работ по созданию системы автоматизированного реферативного перевода авторами были разработаны метод, модель и алгоритм построения денотативной модели текста на основе его внешней формы [6, 7, 8] и, имеющейся в распоряжении системы, модели предметной области [9]. Инфологическая модель представления показана на рис. 1.

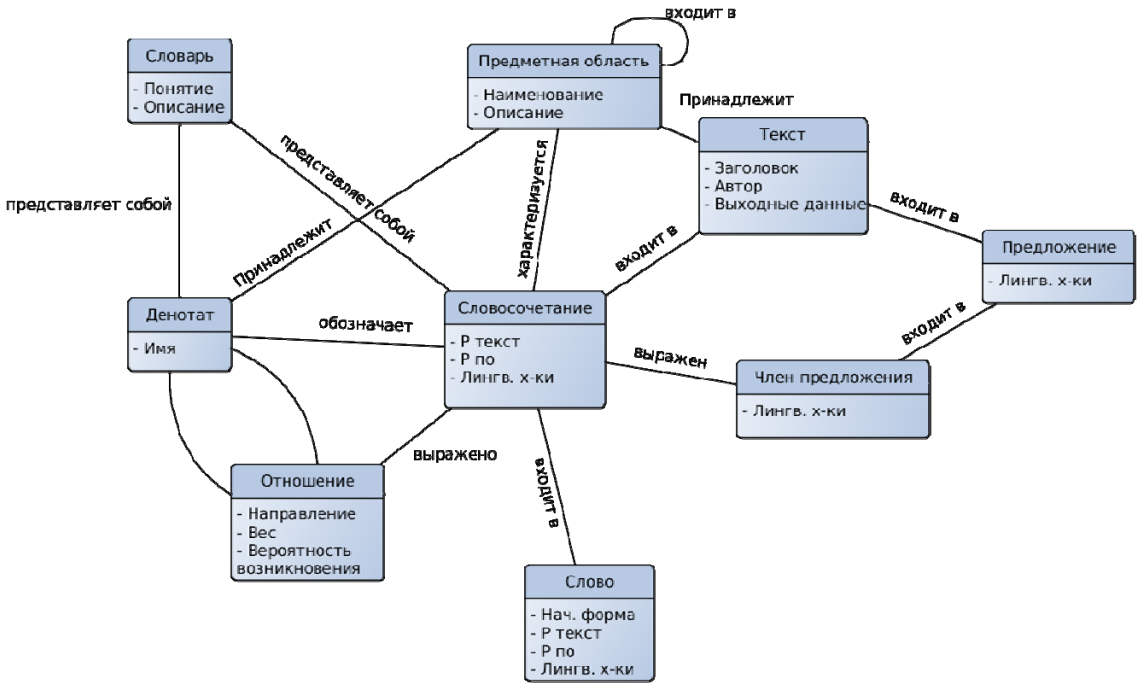

Рис. 1. – Диаграмма «сущность-связь» (фрагмент)

Можно видеть, что модель основного содержания представлена т. н. «денотатными парами» – цепочками «денотат — отношение — денотат». Отношение в этом случае представлено в виде связи с арностью n:m, что отражает тот факт, что в реальном мире так или иначе все понятия как-то связаны друг с другом (напр.: «рыба сидела на дереве»), однако значимость разных связей для понимания текста и их вероятность встречи в конкретном тексте разная. Также имеется атрибут «направление», позволяющий одновременно закодировать такие отношения как «студенты учатся у

преподавателей» и «преподаватели учатся у студентов». Оба утверждения истинны, но для описания, скажем структуры вуза большее значение имеет первый вариант. При анализе текстов именно он должен получить больший вес и вероятность.

Понятие «словосочетание» в данной модели отличается от общепринятого в лингвистике тем, что может содержать и одно слово. Это сделано для универсальности алгоритма обработки входного текста.

Словарь нужен для установления неявных связей между понятиями алгоритмическим путем (за счет нечеткого сравнения словарных статей), а также для возможности расширения текста реферата дополнительными сведениями из него.

Предметная область имеет доменную структуру [10, 11], что позволяет указывать разную вероятность вхождения того или иного понятия в текст в зависимости от контекста. Также в ряде случаев это позволяет разрешать лингвистические неопределенности, свойственные тексту на естественном языке.

Результат разбора текста сохраняется в сущностях «Предложение», «Член предложения» и т. д., что позволяет привязать распознанные денотатные пары к предложениям текста.

Экспериментальные исследования

Эксперименты по обработке текста проводились на базе работы [3], в которой содержатся рефераты научно-технических текстов по тематике «жидкие кристаллы» а также денотатные графы, построенные по ним авторами этой работы. Также в работе представлен т. н. «эталонный граф» (Γ_0), который можно считать денотативной моделью предметной области. Граф построен при участии экспертов в области физики жидких кристаллов.

В таблице 1 приведено краткое содержание эталонного графа, составленного авторами статьи на базе графа из [3]. Вес проставлен (отсутствовал в Γ_0) авторами настоящей статьи.

Структура представлена в формате JSON, что позволило ее обрабатывать системой автореферирования, и загружена в программу. Также, для сравнения была построена визуализация Γ_0 при помощи системы GraphViz [12] (см. рис. 2).

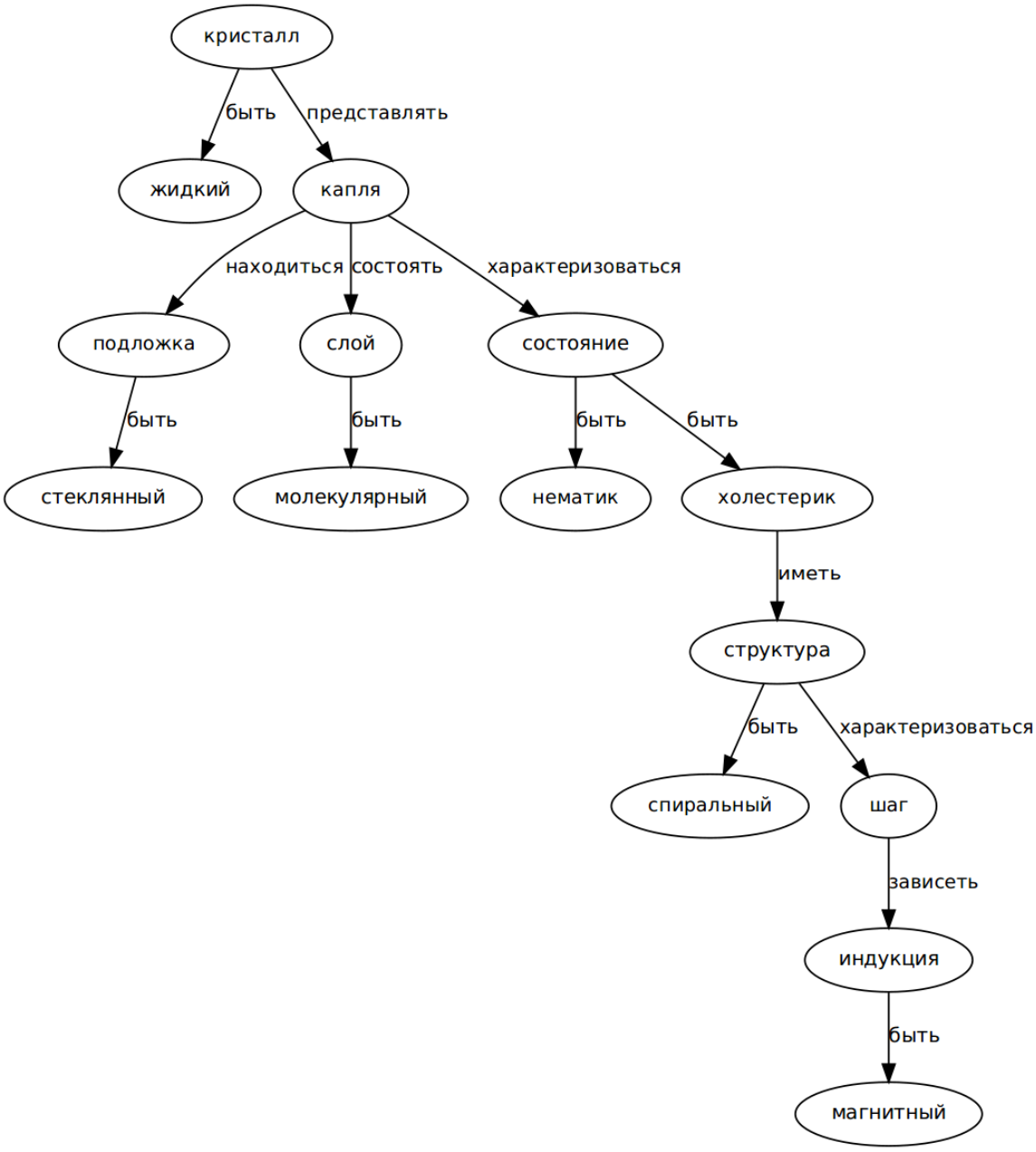

Рис. 2. – Эталонный граф Γ_0

Текст реферата, предложенный системе для анализа представлен на рис. 3. (входной формат системы, JSON).

{
"text": "Жидкий кристалл представлен в виде капли. Капля находится на стеклянной подложке. Капля состоит из молекулярных слоев. Капля ЖК характеризуется состоянием. Состояние может быть нематическим. Состояние может быть холестерическим. Холестерик имеет спиральную структуру. Спиральная структура характеризуется шагом. Шаг зависит от магнитной индукции."
}

Рис. 3. – Текст реферата T_1

Таблица № 1

Описание предметной области в форме денотатных пар

Денотат	Отношение	Денотат	Вес
кристалл	быть	жидкий	0,10
кристалл	представлять	капля	0,10
капля	находиться	подложка	0,20
подложка	быть	стеклянный	0,80
капля	состоять	слой	0,80
слой	быть	молекулярный	0,80
капля	характеризоваться	состояние	0,70
состояние	быть	нематик	0,50
состояние	быть	холестерик	0,50
холестерик	иметь	структура	0,80
структура	быть	спиральный	0,80
структура	характеризоваться	шаг	0,90
шаг	зависеть	индукция	0,60
индукция	быть	магнитный	0,90

В результате анализа текста $Т_1$ система построила денотатный граф $Г_1$, представленный на рис. 4.

Несложно заметить, что $Г_1$ практически идентичен $Г_0$. Это происходит потому, что текст $Т_1$ составлен из ядерных предложений, идентичных денотатным парам, представленным в таблице 1.

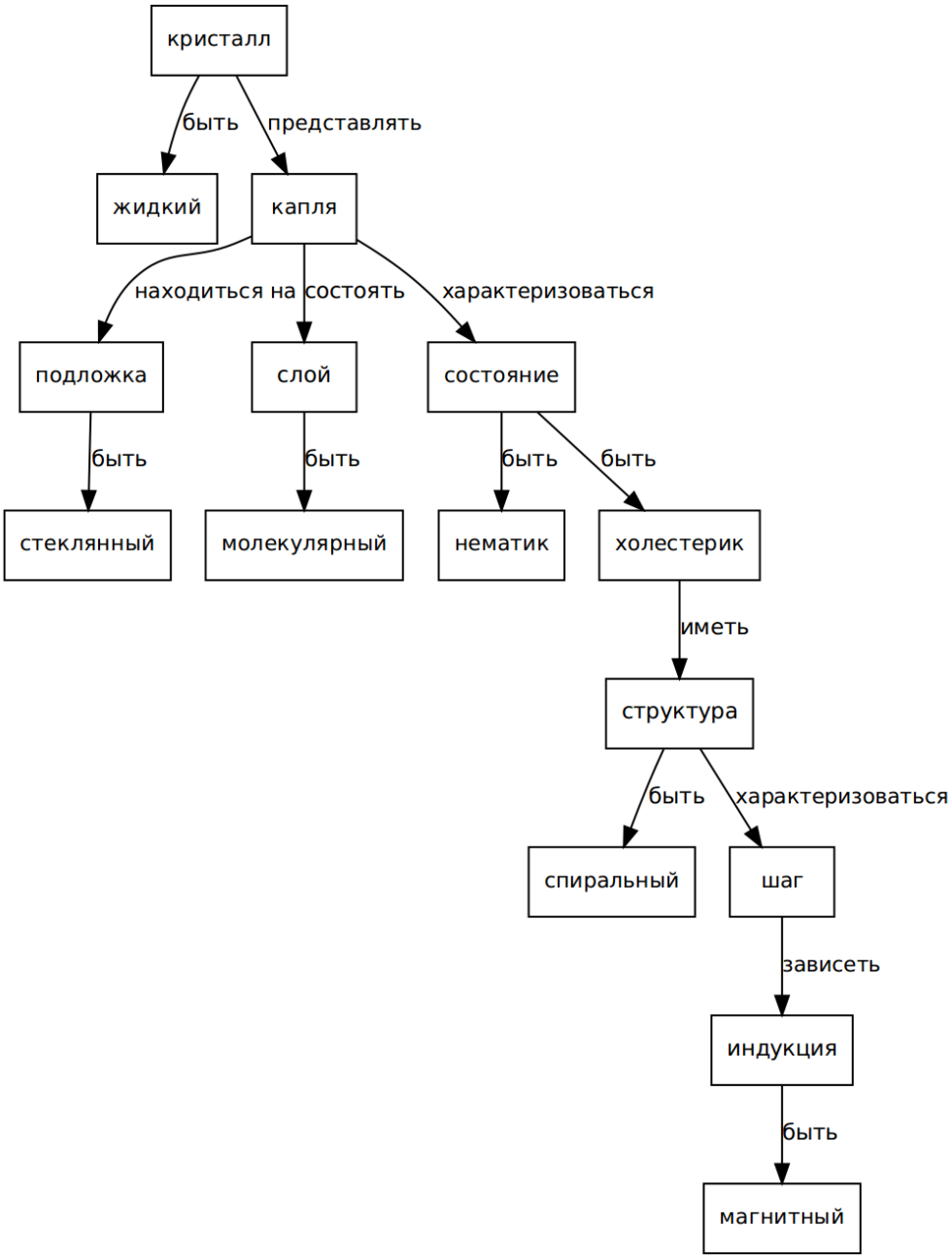

Рис. 4. – Граф $Г_1$, построенный системой по тексту $Т_1$

Более интересные результаты получаются при предъявлении системе текста, в котором имеются отношения, ей неизвестные. Так, например, текст

энциклопедического характера T_2 (см. рис. 5) был проанализирован с явными ошибками (см. рис.6).

{
"text": "Жидкие кристаллы (сокращённо ЖК; англ. liquid crystals, LC) — это фазовое состояние, в которое переходят некоторые вещества при определенных условиях (температура, давление, концентрация в растворе). Жидкие кристаллы обладают одновременно свойствами как жидкостей (текучесть), так и кристаллов (анизотропия). По структуре Жидкие кристаллы представляют собой вязкие жидкости, состоящие из молекул вытянутой или дискообразной формы, определённым образом упорядоченных во всем объёме этой жидкости. Наиболее характерным свойством Жидкие кристаллы является их способность изменять ориентацию молекул под воздействием электрических полей, что открывает широкие возможности для применения их в промышленности. По типу Жидкие кристаллы обычно разделяют на две большие группы: нематики и смектики. В свою очередь нематики подразделяются на собственно нематические и холестерические жидкие кристаллы."
}

Рис. 5. – Текст реферата T_2

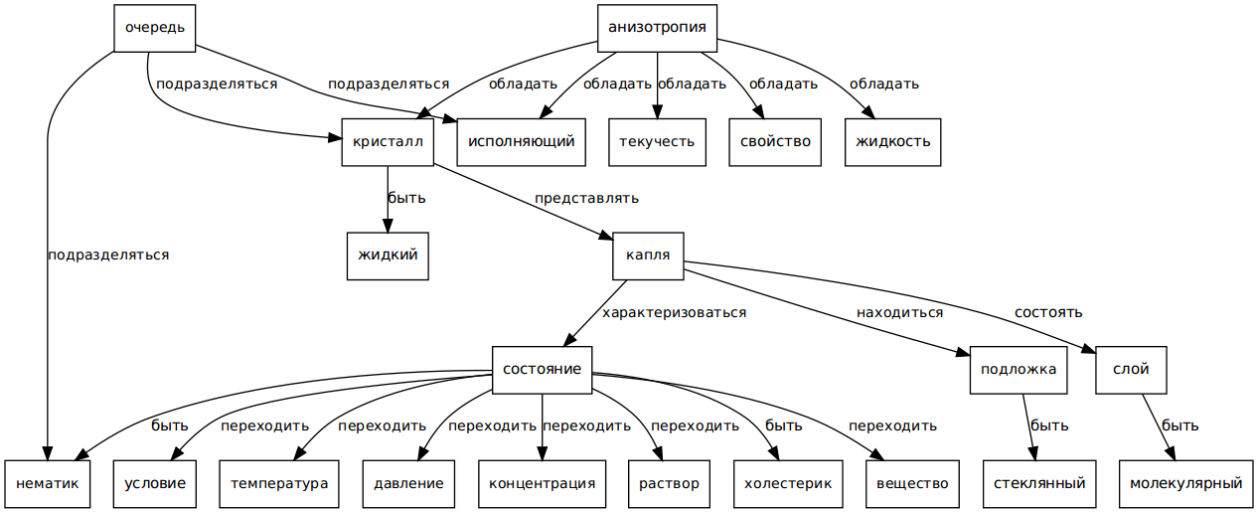

Рис. 6. Граф $Г_2$, построенный по тексту T_2

Основные ошибки, которые можно выделить это:

1) инверсия отношения (анизотропия — обладать — кристалл),

2) «непонимание» оборота «в свою очередь».

Для коррекции возникших ошибок дополним эталонный граф следующими денотатными парами (таблица 2).

Таблица № 2

Дополнения к предметной области

Денотат	Отношение	Денотат	Вес
кристалл	обладать	анизотропия	0,5
нематик	подразделяться	холестерик	0,5
нематик	подразделяться	нематик	0,2

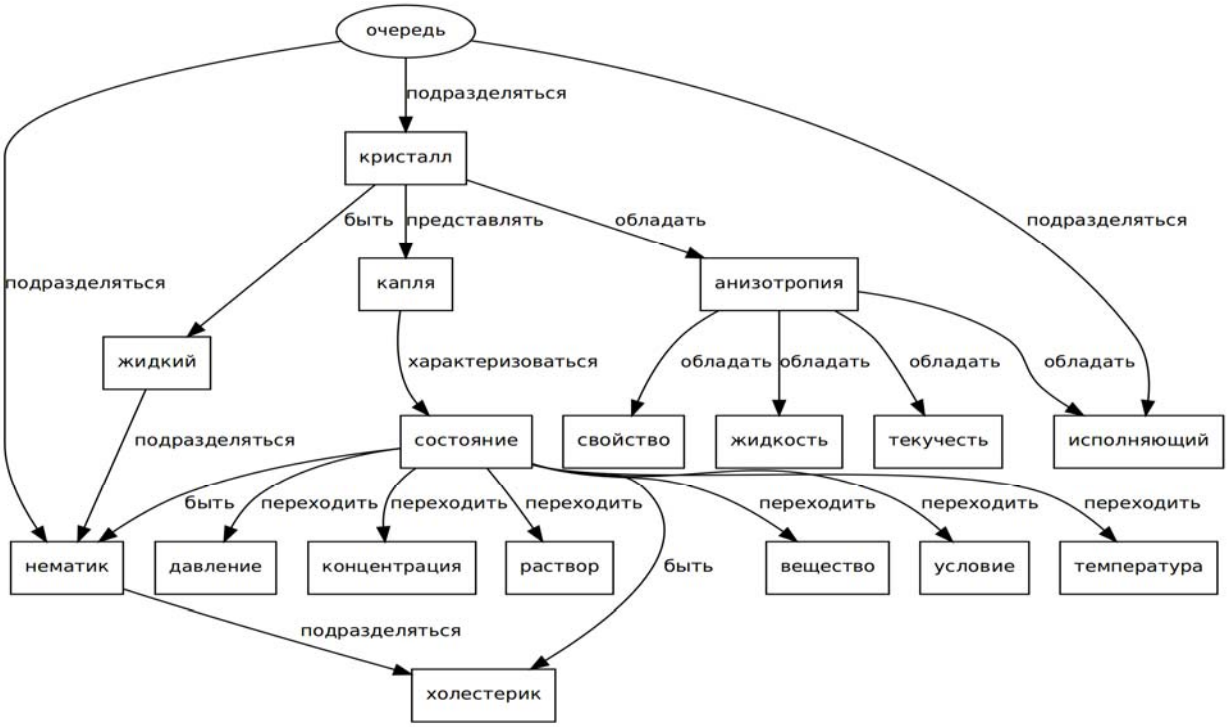

Рис. 7. – Граф $Г_{2\text{-}a}$ по тексту $Т_2$

Как можно видеть (рис. 7), инверсия отношения частично исчезла, утверждение «кристалл — обладать — анизотропией» «пересилило» неправильную интерпретацию грамматической структуры, оборот «в свою очередь» не стал «понятнее» системе, но стал оказывать меньшее влияние на результат (выделен овалом авторами, для наглядности). Можно отметить еще одну ошибку (имеется как в $Г_2$, так и в $Г_{2\text{-}a}$) системы – выделение денотата «исполняющий». Это явление вызвано не вполне корректной работой библиотечного ПО, используемого для получения лингвистических характеристик слов и предложений текста.

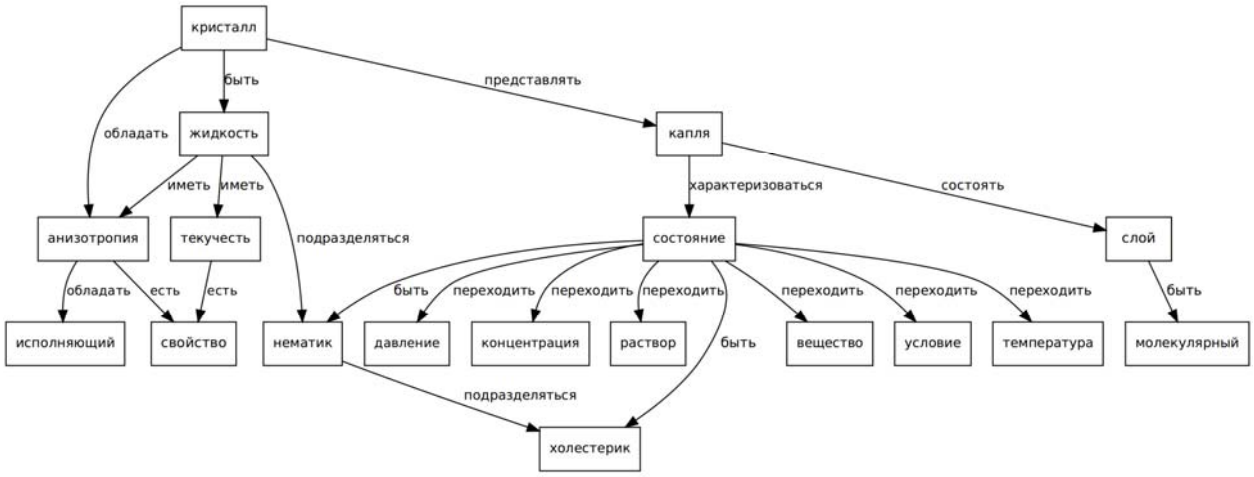

Рис. 8. – Граф $Г_{2\text{-}б}$ по тексту $Т_2$

Далее, в предметную область были внесены следующие утверждения (в виде денотатных пар): «текучесть — есть — свойство — жидкости» и «анизотропия — есть — свойство — свойство — жидкости». Это, как видно из графа $Г_{2\text{-}б}$ (рис. 8) привело к исчезновению инвертированных отношений. «Непонятный» системе оборот «в свою очередь» был исключен из текста.

Выводы

В результате экспериментов установлено, что от того, насколько полно описана предметная область зависит результат интерпретации текста. Также наглядно показано, что когда система не имеет опоры на «знания» о предметной области, она пытается извлекать денотатные пары из грамматической структуры текста, что приводит к ошибочному пониманию текста. Тем не менее полученные рефераты отражают содержание исходного текста (в графовой форме).

В дальнейшем необходимо дополнить систему подсистемами распознавания устоявшихся речевых оборотов типа «в свою очередь», «таким образом» и т.п., которые не влияют на содержание текста. Дальнейшее развитие представленной модели позволит улучшить алгоритмы классификации [13] и индексации документов.

Литература

1. 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).

2. Шепелев А.Н., Букатов А.А., Пыхалов А.В., Березовский А.Н. Анализ подходов и средств обработки сервисных журналов // Инженерный вестник Дона. 2013. №4. URL: ivdon.ru/ru/magazine/archive/n4y2013/1966.

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

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

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

6. Герте Н.А., Нестерова Н.М. Реферирование как способ извлечения и представления основного содержания текста // Вестник Пермского университета. Российская и зарубежная филология. 2013. №4/24. С. 127-132.

7. Герте Н.А. «Эквивалентность» и «адекватность» в реферативном переводе в свете скопос-теории // Межкультурная ↔ интракультурная коммуникация: теория и практика обучения и перевода: материалы III Международной научно-методической конференции. Уфа: РИЦ БашГУ, 2014. С. 109-114.

8. Герте Н.А., Курушин Д.С., Нестерова Н.М. Свертывание информации в процессе реферирования: методы и возможные пути формализации // Вестник ПНИПУ. Проблемы языкознания и педагогики. 2013. №7(49). С. 188-196.

9. Курушин Д.С., Нестерова Н.М., Овчинникова И.Г. О возможном подходе к созданию системы автоматического реферирования // Вопросы психолингвистики. 2014. №2(20). С. 123-127.

10. 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.

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

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

13. 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.

References

1. 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).

2. Shepelev A.N., Bukatov A.A., Pykhalov A.V., Berezovskiy A.N. Inženernyj vestnik Dona (Rus). 2013. №4. URL: ivdon.ru/ru/magazine/archive/n4y2013/1966.

3. 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.

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

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

6. Gerte N.A., Nesterova N.M. Vestnik Permskogo universiteta. Rossiyskaya i zarubezhnaya filologiya. 2013. №4/24. pp. 127-132.

7. Gerte N.A. Mezhkul'turnaya ↔ intrakul'turnaya kommunikatsiya: teoriya i praktika obucheniya i perevoda: materialy III Mezhdunarodnoy nauchno-metodicheskoy konferentsii (Intercultural ↔ intrakulturnaya communication: theory and practice of teaching and translation: Proceedings of the III International Scientific Conference). Ufa: RITs BashGU, 2014. pp. 109-114.

8. Gerte N.A., Kurushin D.S., Nesterova N.M. Vestnik PNIPU. Problemy yazykoznaniya i pedagogiki. 2013. №7 (49). pp. 188-196.

9. Kurushin D.S., Nesterova N.M., Ovchinnikova I.G. Voprosy psikholingvistiki. 2014. №2(20). pp. 123-127.

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

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

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

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