На правах рукописи

Усталов Дмитрий Алексеевич

Модели, методы и алгоритмы построения семантической сети слов для задач обработки естественного языка

Специальность 05.13.17 —
«теоретические основы информатики»

Автореферат
диссертации на соискание ученой степени
кандидата физико-математических наук

Челябинск — 2017

Работа выполнена в отделе вычислительной техники ФГБУН Институт математики и механики им. Н.Н.Красовского Уральского отделения Российской академии наук.

Научный руководитель: **Созыкин Андрей Владимирович**,
кандидат технических наук,
заведующий отделом, ФГБУН Институт математики и механики им. Н.Н.Красовского Уральского отделения Российской академии наук

Официальные оппоненты: **Лукашевич Наталья Валентиновна,**
доктор технических наук,
ведущий научный сотрудник, ФГБОУ ВО «Московский государственный университет имени М.В.Ломоносова»

Турдаков Денис Юрьевич,
кандидат физико-математических наук,
заведующий отделом, ФГБУН Институт системного программирования им. В.П. Иванникова Российской академии наук

Ведущая организация: ФГАОУ ВО «Казанский (Приволжский) федеральный университет»

Защита состоится 21 февраля 2018 г. в 12:00 часов на заседании диссертационного совета Д 212.298.18 при ФГАОУ ВО «Южно-Уральский государственный университет (национальный исследовательский университет)» по адресу: 454080, г. Челябинск, пр. Ленина, 76, ауд. 1001.

С диссертацией можно ознакомиться в библиотеке Южно-Уральского государственного университета и на сайте: https://www.susu.ru/ru/dissertation/d-21229818/ustalov-dmitriy-alekseevich.

Автореферат разослан «___» _____ 20___ г.

Ученый секретарь
диссертационного совета М. Л. Цымблер

Общая характеристика работы

Актуальность темы. Сегодня наблюдается взрывной рост количества информации, создаваемой людьми и машинами на естественном языке. Аналитическое агентство IDC прогнозирует рост совокупного объема данных, накопленных человечеством, до 163 зеттабайт к 2025 году. Основной частью таких данных являются неструктурированные данные, такие как фотографии, видеозаписи, аудиозаписи, а также тексты на естественном языке.

Язык обладает многозначностью, которая проявляется на разных уровнях: от уровня отдельных звуков в устной речи до уровня значения отдельных слов и предложений в письменном тексте. Несмотря на то, что люди хорошо справляются с разрешением многозначности самостоятельно, проблема машинного понимания естественного языка является сложной и требует специальных автоматических методов. Постоянное увеличение интенсивности потока входящей текстовой информации делает все более важной задачу математического моделирования естественного языка, в частности — русского языка.

Важнейшей проблемой является лексическая многозначность, требующая от машины понимания контекста и предметной области, в которой употребляется каждое многозначное слово. Такие сведения представляются в семантических сетях — специальных высококачественных базах знаний, представляющих машиночитаемые сведения об окружающем мире в виде понятий и связей между ними. Связи между понятиями задают семантическую иерархию, которая позволяет решать различные задачи машинного понимания естественного языка и является критически важным элементом семантических сетей. В настоящее время, наиболее известной семантической сетью в области обработки естественного языка является семантическая сеть WordNet для английского языка, связи в которой формируются между синсетами — множествами синонимов.

Семантические сети применяются при решении большого количества важнейших прикладных задач обработки естественного языка. В системах разрешения лексической многозначности и системах машинного перевода, семантические сети представляют известные значения слов заданного языка. В вопросно-ответных системах, таких как IBM Watson, семантические сети задают сведения об объектах предметной области и связях между ними. В системах поиска сущностей, таких как Google Knowledge Graph, семантические сети представляют атрибуты, понятные и людям, и машинам. Высококачественные семантические сети широко используются в качестве золотого стандарта для оценки эффективности систем автоматической обработки естественного языка.

Создание высококачественных баз знаний вручную является длительной и ресурсоемкой задачей, поэтому исследователи уделяют большое внимание вопросу автоматического построения семантических ресурсов, таких как семантические сети. Существующие методы автоматического построения семантических сетей используют высококачественные исходные данные, что затрудняет их применение для автоматической обработки текста на языках,

представляющих другие языковые группы. Например, славянских и балтийских языков. Основное внимание исследователей уделяется английскому языку, для которого сегодня доступно большое количество высококачественных баз знаний и других языковых ресурсов.

Проблема доступности и качества машиночитаемых семантических ресурсов осложняется наличием ошибок или пропущенными данными в существующих словарях. Методы машинного обучения, особенно — методы обучения без учителя, позволяют обнаруживать скрытые закономерности в неструктурированных данных. Применение таких методов может повысить полноту доступных семантических ресурсов. Таким образом, **актуальной** является задача развития методов автоматического построения семантических сетей за счет структурирования и расширения существующих слабоструктурированных словарей, не содержащих сведений о значениях слов.

Цель и задачи исследования. *Целью* данной работы является разработка моделей, методов и алгоритмов построения семантической сети, связывающей лексические значения слов семантическим отношением на основе материалов слабоструктурированных словарей, а также разработка на их основе комплекса программ автоматического построения такой семантической сети.

Для достижения этой цели необходимо было решить следующие *задачи*:

1. Разработать математическую модель представления лексических значений слов и связей между ними в виде семантической сети слов.
2. Разработать метод и алгоритм построения синсетов на основе разрешения многозначности слов.
3. Разработать метод и алгоритм построения и расширения однозначных семантических связей между многозначными словами.
4. Реализовать разработанные модели, методы и алгоритмы в виде комплекса программ, позволяющего построить семантическую сеть слов на основе слабоструктурированных языковых ресурсов.
5. Провести вычислительные эксперименты, подтверждающие эффективность предложенных методов.

Научная новизна работы заключается в следующем:

– разработана оригинальная модель представления значений слов и семантических связей между ними в виде семантической сети слов;
– предложены новый метод и алгоритм построения синсетов путем формирования и кластеризации вспомогательного графа значений слов;
– предложены новый метод и алгоритм построения и расширения однозначных семантических связей между многозначными словами на основе иерархических контекстов;
– разработан комплекс программ автоматического построения семантической сети слов на основе предложенных моделей, методов и алгоритмов.

Теоретическая ценность работы состоит в том, что в ней дано формальное описание методов, алгоритмов и архитектурных решений, позволяющих производить автоматическое построение семантической сети слов на основе

слабоструктурированных языковых ресурсов. **Практическая ценность** работы заключается в том, что на базе разработанных моделей, методов и алгоритмов разработан комплекс программ автоматического построения семантической сети слов, позволяющий повысить полноту сведений о семантических связях. Разработанные методы, алгоритмы и программное обеспечение могут применяться для построения интеллектуальных поисковых систем, систем машинного понимания текста, систем общения, и других информационных систем, основанных на знаниях.

Методология и методы исследования. Методологической основой исследования является теория множеств и теория графов. Для построения синсетов и связывания понятий использовались методы компьютерной лингвистики и машинного обучения. При разработке комплекса программ построения семантической сети слов применялись методы объектно-ориентированного проектирования и язык UML.

Степень достоверности результатов. Все полученные результаты подтверждаются экспериментами, проведенными в соответствии с общепринятыми стандартами.

Апробация работы. Основные положения диссертационной работы, разработанные модели, методы, алгоритмы и результаты вычислительных экспериментов докладывались автором на следующих международных научных конференциях:

- 55-я международная конференция Ассоциации по компьютерной лингвистике (ACL 2017) (30 июля – 4 августа 2017 г., Канада, г. Ванкувер);
- 23-я международная конференция по компьютерной лингвистике «Диалог 2017» (31 мая – 3 июня 2017 г., Москва);
- 15-я международная конференция европейского отделения Ассоциации по компьютерной лингвистике (EACL 2017) (3–7 апреля 2017 г., Испания, г. Валенсия);
- Открытая международная конференция ИСП РАН (1–2 декабря 2016 г., Москва);
- 17-я всероссийская конференция молодых ученых по математическому моделированию и информационным технологиям (30 октября – 3 ноября 2016 г., Новосибирск);
- 5-я международная конференция по анализу изображений, социальных сетей и текстов (АИСТ'2016) (7–9 апреля 2016 г., Екатеринбург);
- 21-я международная конференция по компьютерной лингвистике «Диалог 2015» (27–30 мая 2015 г., Москва);
- 16-я международная суперкомпьютерная конференция «Научный сервис в сети Интернет: многообразие суперкомпьютерных миров» (22–27 сентября 2014 г., Новороссийск);
- 14-я международная конференция европейского отделения Ассоциации по компьютерной лингвистике (EACL 2014) (26–30 апреля 2014 г., Швеция, г. Гетеборг);

– 3-я международная конференция по анализу изображений, социальных сетей и текстов (АИСТ'2014) (10–12 апреля 2014 г., Екатеринбург).

Публикации. По теме диссертации опубликовано восемь печатных работ. Работы [1–4] опубликованы в журналах, включенных ВАК в перечень изданий, в которых должны быть опубликованы основные результаты диссертаций на соискание ученой степени доктора и кандидата наук. Работы [5–7] опубликованы в изданиях, индексируемых в Scopus и Web of Science. В работе [1] научному руководителю Созыкину А. В. принадлежит постановка задачи, Усталову Д. А. — все полученные результаты. В работе [6] постановка задачи принадлежит Биманну К. и Панченко А. И., результаты экспериментов по материалам англоязычных словарей принадлежат Арефьеву Н. В., разработанный метод и результаты экспериментов по материалам русскоязычных словарей принадлежат Усталову Д. А. В работе [8] результаты экспериментов по материалам англоязычных словарей принадлежат Панченко А. И. и Биманну К., все остальные результаты принадлежат Усталову Д. А. В рамках выполнения диссертационной работы получено одно свидетельство Роспатента о государственной регистрации программы для ЭВМ.

Структура и объем работы. Диссертация состоит из введения, четырех глав, заключения и библиографии. В приложении 1 приведены основные обозначения, используемые в диссертации. Приложение 2 содержит список терминов, используемых в диссертации. Объем диссертации составляет 129 страниц, включая 24 рисунка и 9 таблиц. Список литературы содержит 105 наименований.

Содержание работы

Во **введении** обосновывается актуальность исследований, проводимых в рамках данной диссертационной работы, приводится обзор научной литературы по изучаемой проблеме, формулируется цель, ставятся задачи работы, излагается научная новизна и практическая значимость представляемой работы.

Первая глава посвящена обзору работ по автоматизированному построению семантических сетей для решения задач автоматической обработки естественного языка. Перечислены трудности, возникающие при построении семантических сетей. В настоящее время не разработаны методы построения семантической сети путем интеграции существующих слабоструктурированных языковых ресурсов; существующие методы предполагают высокое качество исходных данных.

Вторая глава посвящена разработке модели семантической сети слов, методов и алгоритмов ее автоматического построения. Вводится семантическая сеть слов. Описывается оригинальный метод построения синсетов на основе графа синонимов и приводится соответствующий алгоритм. Описывается оригинальный метод построения и расширения семантических связей между значениями слов на основе иерархических контекстов и приводится соответствующий алгоритм.

Семантическая сеть слов. Пусть словник V — множество всех слов. Пусть $\mathcal{V}$ — множество всех лексических значений слов. Каждому слову $u \in V$ ставится в соответствие множество лексических значений $\text{senses}(u) \subseteq \mathcal{V}$. Пусть $\mathcal{R} \subset \mathcal{V} \times \mathcal{V}$ — асимметричное отношение между лексическими значениями слов; $(w, h) \in \mathcal{R}$ является такой упорядоченной парой, что $w \in \mathcal{V}$ является нижестоящим значением слова по отношению к вышестоящему значению слова $h \in \mathcal{V}$.

***Определение* 1.** *Семантическая сеть слов $\mathcal{N} = (\mathcal{V}, \mathcal{R})$ — это семантическая сеть, понятия которой — лексические значения слов $\mathcal{V}$, а множество дуг $\mathcal{R} \subset \mathcal{V} \times \mathcal{V}$ порождается асимметричным отношением на множестве $\mathcal{V}$.*

Для построения семантической сети слов предлагается модифицированный метод ECO (англ. *Extraction, Clustering, Ontologisation*, рис. 1), отличающийся тем, что на этапе *извлечения* наряду со словарями извлекаются также и векторные представления слов, на этапе *кластеризации* производится построение множества *понятий* $\mathcal{V}$ и синсетов $\mathcal{S}$, на этапе *связывания* производится построение, расширение и построение *связей* между понятиями при помощи иерархических контекстов построенных синсетов.

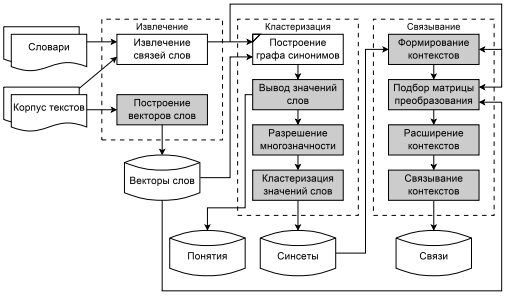

Рис. 1 — Общая схема предлагаемого метода построения семантической сети слов: в блоки исходного метода ECO, помеченные уголком сверху слева, внесены изменения; новые блоки выделены цветом

***Определение* 2.** *Синсет $S \in \mathcal{S}$ — это множество $S \subseteq \mathcal{V}$, такое, что все пары элементов S принадлежат отношению синонимии.*

Метод построения синсетов. Пусть словарь синонимов $D \subseteq V \times V$ — это отношение синонимии между словами; пусть задана некоторая мера семантической близости слов $\text{sim}_{\text{word}} : (u, v) \to \mathbb{R}, \forall u \in V, v \in V$.

***Определение* 3.** *Граф синонимов $W = (V, E)$ — это неориентированный взвешенный граф, множество вершин V которого является словником, а множество ребер E порождается отношением синонимии на словнике.*

Каждое ребро графа W взвешивается с использованием меры sim_{word}. На основе допущения о том, что связанные скопления вершин в графе синонимов обозначают одно и то же понятие, множество понятий семантической сети слов $\mathcal{V}$ и множество синсетов $\mathcal{S}$ можно построить путем кластеризации графа W. В свою очередь, кластеризация графа W затруднена тем, что данный граф содержит как однозначные, так и многозначные слова. Таким образом, предлагается

производить кластеризацию вспомогательного графа значений слов $\mathcal{W} = (\mathcal{V}, \mathcal{E})$, где $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$. Используется следующая *постановка задачи* построения синсетов: найти все синсеты $\mathcal{S}$ в графе $\mathcal{W}$ такие, что в каждом синсете $S \in \mathcal{S}$ любая пара значений слов $a \in S, b \in S$ находится в отношении синонимии.

***Определение* 4.** *Граф значений слов $\mathcal{W} = (\mathcal{V}, \mathcal{E})$ — это неориентированный взвешенный граф, множество вершин которого состоит из лексических значений слов, а множество ребер порождается отношением синонимии на множестве лексических значений слов.*

Вывод значений слов производится путем кластеризации окрестности W_u каждой вершины $u \in V$ в графе W при помощи какого-либо метода жесткой кластеризации графа. Каждый i-й полученный кластер записывается в контекст $\text{ctx}(u^i)$ i-го значения слова u, который представляется в виде «мешка слов». Множество понятий семантической сети слов является объединением всех множеств значений слов: $\mathcal{V} = \cup_{u \in V} \text{senses}(u)$.

***Определение* 5.** *Контекст $\text{ctx}(u^i)$ — множество синонимов слова $u \in V$ в значении под номером $1 \leq i \leq |\,\text{senses}(u)|$.*

Пусть задана некоторая мера близости контекстов $\text{sim}_{\text{ctx}} : (\text{ctx}(a), \text{ctx}(b)) \to \mathbb{R}, \forall a \in \mathcal{V}, b \in \mathcal{V}$. Поскольку элементами контекстов являются слова без указания значений, производится разрешение многозначности контекста каждого значения слова $s \in \mathcal{V}$. Каждому элементу $u \in \text{ctx}(s)$ ставится в соответствие значение $\hat{u} \in \mathcal{V}$ с наиболее близким контекстом: $\hat{u} \in \arg\max_{u' \in \text{senses}(u)} \text{sim}_{\text{ctx}}(\text{ctx}(s), \text{ctx}(u'))$. Производится построение контекстов с разрешенной многозначностью для каждого $s \in \mathcal{V}$: $\widehat{\text{ctx}}(s) = \{\hat{u} : u \in \text{ctx}(s)\}$. На основе контекстов со снятой многозначностью формируется множество ребер $\mathcal{E}$ графа значений слов $\mathcal{W}$: $\mathcal{E} = \{\{\hat{u}, \hat{v}\} \in \mathcal{V} \times \mathcal{V} : \hat{v} \in \widehat{\text{ctx}}(\hat{u})\}$. В качестве заключительного шага производится кластеризация графа $\mathcal{W}$. Полученное в результате жесткой кластеризации графа $\mathcal{W}$ множество кластеров $\mathcal{S}$ является искомым множеством синсетов.

На основе метода построения синсетов предложен алгоритм Watset. Входными данными для алгоритма является словарь синонимов $D \subseteq V \times V$. Результатом работы алгоритма является множество значений слов $\mathcal{V}$ и множество синсетов $\mathcal{S}$. Алгоритм имеет четыре гиперпараметра:

- $\text{Cluster}_{\text{Local}}$ — алгоритм жесткой кластеризации графа, используемый для кластеризации окрестностей вершин в графе синонимов при выводе лексических значений слов;
- $\text{Cluster}_{\text{Global}}$ — алгоритм жесткой кластеризации графа, используемый для поиска синсетов в графе значений слов;
- $\text{sim}_{\text{word}} : (u, v) \to \mathbb{R}$ — мера близости слов $u \in V$ и $v \in V$;
- $\text{sim}_{\text{ctx}} : (\text{ctx}(a), \text{ctx}(b)) \to \mathbb{R}$ — мера близости контекстов значений слов $a \in \mathcal{V}$ и $b \in \mathcal{V}$.

Алгоритм Watset состоит из головной процедуры и трех вспомогательных процедур построения графа синонимов, вывода значений заданного слова и разрешения многозначности контекстов слова.

Головная процедура. В общем виде, головная процедура выглядит следующим образом:

Шаг 1. **Построить граф синонимов** W;
Шаг 2. Для всех слов $u \in V$ выполнить цикл
Шаг 2.1. **Произвести вывод значений слова** u;
Шаг 3. Конец цикла;
Шаг 4. Построить множество значений всех слов: $\mathcal{V} \leftarrow \bigcup_{u \in V} \text{senses}(u)$;
Шаг 5. Для всех значений слов $s \in \mathcal{V}$ выполнить цикл.
Шаг 5.1. **Разрешить многозначность контекста** s;
Шаг 6. Конец цикла;
Шаг 7. Построить множество ребер: $\mathcal{E} \leftarrow \{\{\hat{u}, \hat{v}\} \in \mathcal{V} \times \mathcal{V} : \hat{v} \in \widehat{\text{ctx}}(\hat{u})\}$;
Шаг 8. Выполнить кластеризацию графа $\mathcal{W} = (\mathcal{V}, \mathcal{E})$:
$\mathcal{S} \leftarrow \text{Cluster}_{\text{Global}}(\mathcal{W})$;
Шаг 9. Стоп.

Процедура построения графа синонимов. Входными данными для процедуры является словарь синонимов D. Результатом выполнения процедуры является граф синонимов $W = (V, E)$, взвешенный при помощи меры sim_{word}. Процедура выглядит следующим образом:

Шаг 1.1. $V \leftarrow \bigcup_{(u,v) \in D} \{u, v\}$;
Шаг 1.2. $E \leftarrow \{\{u, v\} \in V \times V : (u, v) \in D, u \neq v\}$;
Шаг 1.3. Для всех ребер $\{u, v\} \in E$ выполнить цикл
Шаг 1.3.1. $\text{weight}(u, v) \leftarrow \text{sim}_{\text{word}}(u, v)$;
Шаг 1.4. Конец цикла;
Шаг 1.5. Конец процедуры.

Процедура вывода значений слова. Входными данными для процедуры является граф синонимов $W = (V, E)$ и заданное слово $u \in V$. Результатом выполнения процедуры является множество $\text{senses}(u)$, содержащее все обнаруженные значения слова u, причем для каждого обнаруженного значения составлен контекст, представляющий синонимы слова в данном значении. Процедура выглядит следующим образом:

Шаг 2.1.1. $\text{senses}(u) \leftarrow \varnothing$;
Шаг 2.1.2. Извлечь вершины окрестности вершины u:
$V_u \leftarrow \{v \in V : \{u, v\} \in E\}$;
Шаг 2.1.3. Извлечь ребра окрестности вершины u:
$E_u \leftarrow \{\{v, w\} \in E : v \in V_u, w \in V_u\}$;
Шаг 2.1.4. Выполнить кластеризацию графа $W_u = (V_u, E_u)$:
$C \leftarrow \text{Cluster}_{\text{Local}}(W_u)$;
Шаг 2.1.5. $i \leftarrow 1$;
Шаг 2.1.6. $\text{ctx}(u^i) \leftarrow C_i$;

Шаг 2.1.7. $\text{senses}(u) \leftarrow \text{senses}(u) \cup \{u^i\}$;
Шаг 2.1.8. Если $i < |C|$, то $i \leftarrow i + 1$ и перейти на шаг 2.1.6;
Шаг 2.1.9. Конец процедуры.

Процедура разрешения многозначности контекста. Входными данными для процедуры является заданное значение слова $s \in \mathcal{V}$. Результатом выполнения процедуры является контекст с разрешенной многозначностью $\widehat{\text{ctx}}(s)$. Процедура выглядит следующим образом:
Шаг 6.1.1. $\widehat{\text{ctx}}(s) \leftarrow \varnothing$;
Шаг 6.1.2. Для каждого слова в контексте $u \in \text{ctx}(s)$ выполнить цикл
Шаг 6.1.2.1. $\hat{u} \leftarrow \arg\max_{u' \in \text{senses}(u)} \text{sim}_{\text{ctx}}(\text{ctx}(s), \text{ctx}(u'))$;
Шаг 6.1.2.2. $\widehat{\text{ctx}}(s) \leftarrow \widehat{\text{ctx}}(s) \cup \{\hat{u}\}$;
Шаг 6.1.3. Конец цикла;
Шаг 6.1.4. Конец процедуры.

Сформулирована и доказана следующая теорема.

Теорема. *Пусть* $\deg_{\max}$ *— максимальная степень вершины графа* $W = (V, E)$. *Тогда вычислительная сложность процедуры разрешения многозначности контекста всех значений слов составляет* $O(|V| \deg^4_{\max})$ *при использовании косинусной меры близости контекстов значений слов.*

Метод построения связей. Пусть $R \subset V \times V$ — асимметричное отношение, определенное на словнике. Пусть $(w, h) \in R$ является такой упорядоченной парой, что $w \in V$ является нижестоящим словом по отношению к вышестоящему слову $h \in V$. Используется следующая *постановка задачи* построения связей: для каждого синсета $S \in \mathcal{S}$ найти множество вышестоящих значений слов $\widehat{\text{hctx}}(S) \subset \mathcal{V}$, такое, что каждый элемент $\hat{h} \in \widehat{\text{hctx}}(S)$ является вышестоящим значением по отношению к каждому элементу $s \in S$.

***Определение* 6.** *Иерархический контекст* $\text{hctx}(S) \subset V$ *синсета* $S \in \mathcal{S}$ *— это объединение множеств вышестоящих слов для каждого слова синсета* S.

Пусть $\text{words}(S) \subseteq V$ — множество слов, значения которых включены в синсет S. Тогда каждому синсету $S \in \mathcal{S}$ ставится в соответствие иерархический контекст $\text{hctx}(S) = \{h \in V : (w, h) \in R, w \in \text{words}(S), h \notin \text{words}(S)\}$, который представляется в виде «мешка слов». Поскольку значимость слов в иерархических контекстах различается, предлагается использовать меру tf–idf для взвешивания элементов контекстов: $\text{tf–idf}(h, S, \mathcal{S}) = \text{tf}(h, S) \times \text{idf}(h, \mathcal{S})$, где

$$\text{tf}(h, S) = \frac{|h' \in \text{hctx}(S) : h = h'|}{|\,\text{hctx}(S)|}, \quad \text{idf}(h, \mathcal{S}) = \log \frac{|\mathcal{S}|}{|S' \in \mathcal{S} : h \in \text{hctx}(S')|}.$$

На практике, доступность данных для построения отношения R низка, для чего производится расширение иерархических контекстов. Пусть $\text{NN}_n(\vec{h}) \in V$ — операция поиска $n \in \mathbb{Z}^+$ слов, векторные представления которых соответствуют ближайшим соседям векторного представления $\vec{h}$ слова $h \in \text{hctx}(S)$. Пусть Φ^* — такая матрица, что $\Phi^* \vec{w} = \vec{h}, \forall (w, h) \in R$. Расширение производится путем построения множества кандидатов $M_S = \bigcup_{h \in \text{hctx}(S)} \text{NN}_n(\vec{h}) \setminus \text{hctx}(S)$

и проверки каждого кандидата $h \in M_S$. При выполнении условия $\exists w \in \text{words}(S) : \|\Phi^* \vec{w} - \vec{h}\| < \delta$, где $\delta \in \mathbb{R}^+$ — некоторое пороговое значение, слово-кандидат h включается в иерархический контекст $\text{hctx}(S)$.

Подбор значений элементов матрицы линейного преобразования Φ^* производится методом наименьших квадратов. С целью использования информации об асимметричных связях между словами, предлагается ввести в функцию потерь член стабилизации, влияние которого определяется коэффициентом $\lambda \in \mathbb{R}$:

$$\Phi^* \in \arg\min_{\Phi} \frac{1}{|R|} \left(\sum_{(\vec{w},\vec{h}) \in R} \left\| \Phi \vec{w} - \vec{h} \right\|^2 + \lambda \sum_{(\vec{w},\vec{h}) \in R} ((\Phi^2 \vec{w})^T \vec{w})^2 \right).$$

Пусть задана некоторая мера близости иерархического контекста и слов синсета $\text{sim}_{\text{hctx}} : (\text{hctx}(A), \text{words}(B)) \to \mathbb{R}, \forall A \in \mathcal{S}, B \in \mathcal{S}$. Каждому элементу $h \in \text{hctx}(S)$ ставится в соответствие значение $\hat{h} \in \mathcal{V}$, являющееся элементом синсета, наиболее близкого к $\text{hctx}(S)$: $\hat{h} \in \arg\max_{h' \in \text{senses}(h) : S' \in \mathcal{S}, h' \in S', S \neq S'} \text{sim}_{\text{hctx}}(\text{hctx}(S), \text{words}(S'))$. Затем производится построение иерархического контекста с разрешенной многозначностью для каждого синсета $S \in \mathcal{S}$: $\widehat{\text{hctx}}(S) = \{\hat{h} : h \in \text{hctx}(S)\}$. На основе контекстов со снятой многозначностью формируется множество дуг $\mathcal{R}$ семантической сети слов $\mathcal{N}$: $\mathcal{R} = \bigcup_{S \in \mathcal{S}} S \times \widehat{\text{hctx}}(S)$.

Семантическая сеть слов формируется на основе построенного множества понятий $\mathcal{V}$ и построенного множества связей $\mathcal{R}$. Пример семантической сети слов для многозначного слова «программа» приведен на рис. 2. Слова с различными значениями, но с совпадающими лексемами, не имеют общих связей.

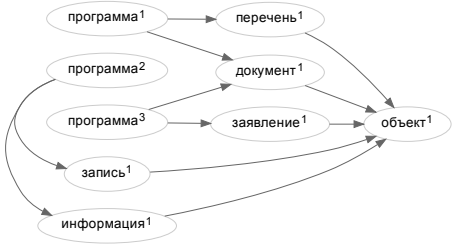

Рис. 2 — Пример фрагмента семантической сети слов

На основе метода построения связей предложен алгоритм Watlink. Входными данными для алгоритма является множество синсетов $\mathcal{S}$ и отношение R. Результатом работы алгоритма является семантическая сеть слов $\mathcal{N}$. Алгоритм имеет пять гиперпараметров:

- $n \in \mathbb{Z}^+$ — количество ближайших соседей, возвращаемых при расширении контекстов;
- $k \in \mathbb{N}$ — количество подпространств при подборе матрицы линейного преобразования;
- $\lambda \in \mathbb{R}$ — влияние стабилизации на функцию потерь при подборе матрицы линейного преобразования;
- $\delta \in \mathbb{R}^+$ — максимальное расстояние до ближайшего соседа, включаемого при расширении иерархического контекста;
- $\text{sim}_{\text{hctx}} : (\text{hctx}(S), S') \to \mathbb{R}$ — мера близости иерархического контекста $\text{hctx}(S) \subseteq V$ и слов синсета $S' \in \mathcal{S}$: $\text{words}(S') \subseteq V$.

Алгоритм Watlink состоит из головной процедуры и трех вспомогательных процедур подбора матриц линейного преобразования, построения иерархического контекста синсета, разрешения многозначности иерархического контекста.

Головная процедура. В общем виде, головная процедура выглядит следующим образом:

Шаг 1. **Подобрать матрицы линейного преобразования**;
Шаг 2. Для всех синсетов $S \in \mathcal{S}$ выполнить цикл
Шаг 2.1. **Построить иерархический контекст синсета** S;
Шаг 3. Конец цикла;
Шаг 4. Для всех синсетов $S \in \mathcal{S}$ выполнить цикл
Шаг 4.1. $\text{tf–idf}(h, S, \mathcal{S}) \leftarrow \text{tf}(h, S) \times \text{idf}(h, \mathcal{S})$;
Шаг 5. Конец цикла;
Шаг 6. Для всех синсетов $S \in \mathcal{S}$ выполнить цикл
Шаг 6.1. **Разрешить многозначность иерархического контекста** $\text{hctx}(S)$;
Шаг 7. Конец цикла;
Шаг 8. Построить связи между значениями слов: $\mathcal{R} \leftarrow \bigcup_{S \in \mathcal{S}} S \times \widehat{\text{hctx}}(S)$;
Шаг 9. Построить семантическую сеть слов $\mathcal{N} \leftarrow (\mathcal{V}, \mathcal{R})$;
Шаг 10. Стоп.

Процедура подбора матриц линейного преобразования. Входными данными для процедуры является $k \in \mathbb{N}$ — количество линейных подпространств, R — множество семантических отношений между словами, $\lambda \in \mathbb{R}$ — важность члена стабилизации. Результатом выполнения процедуры являются k матриц $\Phi_i^* : 1 \leq i \leq k$. Процедура выглядит следующим образом:

Шаг 1.1. Для каждой пары слов $(w, h) \in R$ выполнить цикл
Шаг 1.1.1. $\text{offsets}(w, h) \leftarrow (\vec{h} - \vec{w})$;
Шаг 1.2. Конец цикла;
Шаг 1.3. $C \leftarrow \text{k-means}(\text{offsets}, k)$;
Шаг 1.4. $i \leftarrow 1$;
Шаг 1.5. $\Phi_i^* \leftarrow \arg\min_{\Phi_i} \frac{1}{|R_i|} \sum_{(\vec{w}, \vec{h}) \in R_i} (\|\Phi_i \vec{w} - \vec{h}\|^2 + \lambda((\Phi_i^2 \vec{w})^T \vec{w})^2)$;
Шаг 1.6. Если $i < k$, то $i \leftarrow i + 1$ и перейти на шаг 1.4;
Шаг 1.7. Конец процедуры.

Процедура построения иерархического контекста. Входными данными для процедуры является синсет $S \in \mathcal{S}$ и k матриц линейного преобразования. Результатом выполнения процедуры является иерархический контекст $\text{hctx}(S)$. Процедура выглядит следующим образом:

Шаг 2.1.1. $\text{hctx}(S) \leftarrow \{h \in V : (w, h) \in R, w \in \text{words}(S), h \notin \text{words}(S)\}$;
Шаг 2.1.2. Сформировать множество слов-кандидатов в иерархический контекст: $M_S \leftarrow \bigcup_{h \in \text{hctx}(S)} \text{NN}_n(\vec{h}) \setminus \text{hctx}(S)$;
Шаг 2.1.2. Для каждой пары слов $(w, h) \in \text{words}(S) \times M_S$ выполнить цикл
Шаг 2.1.2.1. Выбрать одну из k матриц линейного преобразования Φ^* для пары слов (w, h) на основе смещения $(\vec{h} - \vec{w})$;

Шаг 2.1.2.2. Если $\|\vec{w}\Phi^* - \vec{h}\| < \delta$, то $\text{hctx}(S) \leftarrow \text{hctx}(S) \cup \{h\}$;
Шаг 2.1.3. Конец цикла;
Шаг 2.1.4. Конец процедуры.

Процедура разрешения многозначности иерархического контекста. Входными данными для процедуры является синсет $S \in \mathcal{S}$ и его иерархический контекст $\text{hctx}(S)$. Результатом выполнения процедуры является иерархический контекст синсета S с разрешенной многозначностью $\widehat{\text{hctx}}(S)$. Процедура выглядит следующим образом:

Шаг 6.1.1. $\widehat{\text{hctx}}(S) \leftarrow \varnothing$;
Шаг 6.1.2. Для каждого слова $h \in \text{hctx}(S)$ выполнить цикл
Шаг 6.1.2.1. $\hat{h} \leftarrow \arg\max_{h' \in \text{senses}(h): S' \in \mathcal{S}, h' \in S', S \neq S'} \text{sim}_{\text{hctx}}(\text{hctx}(S), \text{words}(S'))$;
Шаг 6.1.2.2. $\widehat{\text{hctx}}(S) \leftarrow \widehat{\text{hctx}}(S) \cup \{\hat{h}\}$;
Шаг 6.1.3. Конец цикла;
Шаг 6.1.4. Конец процедуры.

Третья глава посвящена разработке архитектуры комплекса программ, реализующего предложенные модели, методы и алгоритмы. На основе предложенной архитектуры (рис. 3) реализован комплекс программ с использованием языков программирования Python, AWK, Java. Использованы внешние библиотеки scikit-learn, Gensim, TensorFlow и Raptor. Результат работы методов представляется в формализме RDF в виде троек «субъект–предикат–объект» с использованием моделей SKOS и Lemon. При реализации комплекса программ SWN использованы языки программирования Python, AWK и Bash. Исходные тексты разработанных программ доступны в сети Интернет по адресу https://github.com/dustalov/watset, https://github.com/dustalov/projlearn и https://github.com/dustalov/watlink.

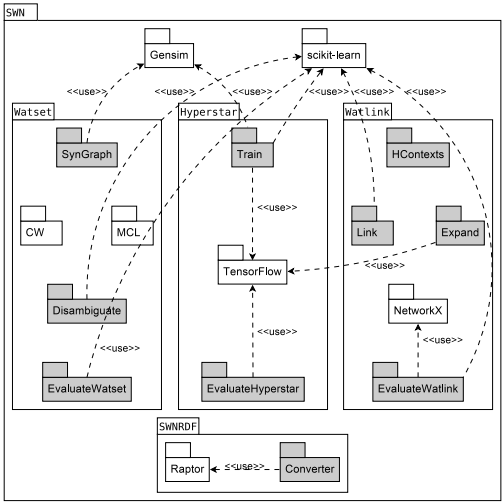

Рис. 3 — UML-диаграмма пакетов; цветом выделены программы, разработанные в рамках данной диссертационной работы

Четвертая глава посвящена проверке адекватности разработанных методов на основе сравнения полученных результатов с результатами, полученными путем использования методов, опубликованных в открытой литературе.

Результаты экспериментальной оценки метода Watset на основе попарной точности, полноты и F_1-меры приведены в табл. 1 по материалам двух различных золотых стандартов: RuWordNet и Yet Another RussNet. Метод Watset получил лучшие значения полноты и F_1-меры на обоих наборах данных. Запись Watset[$\text{Cluster}_{\text{Local}}$, $\text{Cluster}_{\text{Global}}$] означает, что для вывода значений слов использован алгоритм $\text{Cluster}_{\text{Local}}$, а для кластеризации графа значений слов использован алгоритм $\text{Cluster}_{\text{Global}}$.

Таблица 1 — Сравнение методов построения синсетов по материалам RuWordNet и Yet Another RussNet

Метод	# синсетов	# пар	RuWordNet			Yet Another RussNet		
			Точность	Полнота	F_1-мера	Точность	Полнота	F_1-мера
Watset[CW_{nolog}, MCL]	55 369	332 727	0,120	**0,349**	**0,178**	0,402	**0,463**	**0,430**
Watset[MCL, MCL]	36 217	403 068	0,111	0,341	0,168	0,405	0,455	**0,428**
Watset[CW_{top}, CW_{log}]	55 319	341 043	0,116	**0,351**	0,174	0,386	**0,474**	**0,425**
MCL	21 973	353 848	0,155	0,291	**0,203**	0,550	0,340	0,420
Watset[MCL, CW_{top}]	34 702	473 135	0,097	**0,361**	0,153	0,351	**0,496**	0,411
CW_{nolog}	19 124	672 076	0,087	0,342	0,139	0,364	0,451	0,403
MaxMax	27 011	461 748	**0,176**	0,261	**0,210**	**0,582**	0,195	0,292
$\text{CPM}_{k=3}$	4 000	45 231	**0,234**	0,072	0,111	**0,626**	0,060	0,110
ECO	67 645	18 362	**0,724**	0,034	0,066	**0,904**	0,002	0,004

Результаты сравнения методов подбора матрицы линейного преобразования с использованием меры качества hit@10 по тестовой выборке представлены в табл. 2; лучшие значения выделены полужирным начертанием. Значения критериев hit@1 и hit@5 приведены в иллюстративных целях. Выбор лучшего метода и подбор гиперпараметров для оценки метода Watlink производится на основании значения hit@10.

Таблица 2 — Сравнение методов подбора матрицы линейного преобразования по тестовой выборке

Метод	k	hit@1	hit@5	hit@10
Базовый	1	0,0473	0,1095	0,1297
Стабилизированный	1	**0,0522**	**0,1199**	**0,1403**
Базовый	20	0,2090	0,3031	0,3232
Стабилизированный	20	**0,2119**	**0,3120**	**0,3343**

Результаты экспериментальной оценки метода Watlink на основе проверки существования путей в графе по точности, полноты и F_1-меры приведены в табл. 3 по материалам золотого стандарта: тезауруса RuWordNet. Метод Watlink с использованием расширения получил лучшие значения полноты и F_1-меры на этом наборе данных. При использовании метода Watlink применяется обозначение «+ РЛП + ССС» (расширение при помощи линейного преобразования, семантическая сеть слов).

Результаты экспериментов показывают и подтверждают, что предложенные в данной работе методы, модели и алгоритмы позволяют эффективно строить семантическую сеть слов.

Таблица 3 — Сравнение методов построения отношений по материалам RuWordNet

Метод	# связей	Точность	Полнота	F_1-мера
Шаблоны	1 597 651	0,1611	**0,3255**	**0,2155**
Шаблоны + Ч	10 458	**0,3773**	0,0157	0,0302
Шаблоны + Ч + РЛП	10 715	0,3760	0,0160	0,0307
Шаблоны + Ч + РЛП + ССС	47 387	0,1129	0,0722	0,0881
Викисловарь	108 985	**0,3877**	0,0898	0,1458
Викисловарь + РЛП	110 329	**0,3874**	0,0907	0,1469
Викисловарь + РЛП + ССС	179 623	0,1844	**0,3464**	**0,2407**
МАС	36 800	0,1823	0,1502	0,1647
МАС + РЛП	37 702	0,1825	0,1515	0,1655
МАС + РЛП + ССС	99 678	0,1385	0,1883	0,1596
Все словари	149 195	0,1719	0,2590	0,2067
Все словари + РЛП	151 150	0,1720	0,2594	0,2069
Все словари + РЛП + ССС	218 290	0,1687	**0,3867**	**0,2350**

В **заключении** в краткой форме излагаются итоги выполненного диссертационного исследования, представляются отличия диссертационной работы от ранее выполненных родственных работ других авторов, даются рекомендации по использованию полученных результатов и рассматриваются перспективы дальнейшего развития темы.

Основные результаты диссертационной работы

На защиту выносятся следующие новые научные результаты:

1. Предложена модель семантической сети слов, связывающей лексические значения слов семантическим отношением.
2. Разработан метод и алгоритм построения синсетов путем формирования и кластеризации вспомогательного графа значений слов.
3. Разработан метод и алгоритм построения и расширения однозначных семантических связей между многозначными словами.
4. Выполнена реализация комплекса программ автоматического построения семантической сети слов.
5. Проведены вычислительные эксперименты, подтверждающие высокую эффективность разработанных моделей, методов и алгоритмов.

Публикации по теме диссертации

Статьи в журналах из перечня ВАК

1. *Усталов Д., Созыкин А.* Комплекс программ автоматического построения семантической сети слов // *Вестник ЮУрГУ. Серия: Вычислительная математика и информатика*. 2017. Т. 6, № 2. С. 69–83.
2. *Усталов Д.* Семантические сети и обработка естественного языка // *Открытые системы. СУБД*. 2017. № 2. С. 46–47.
3. *Усталов Д.* Обнаружение понятий в графе синонимов // *Вычислительные технологии*. 2017. Т. 22, Спецвып. 1. С. 99–112.

4. *Ustalov D.* Joining Dictionaries and Word Embeddings for Ontology Induction // *Proceedings of the Institute for System Programming*. 2016. Vol. 28, no 6. P. 197–206.

Статьи в изданиях, индексируемых в Scopus и Web of Science

5. *Ustalov D.* Expanding Hierarchical Contexts for Constructing a Semantic Word Network // Computational Linguistics and Intellectual Technologies: Papers from the Annual conference "Dialogue". Volume 1 of 2. Computational Linguistics: Practical Applications, May 31 – June 3, 2017, Moscow, Russia. Moscow, Russia: RSUH, 2017. P. 369–381.
6. *Ustalov D., Arefyev N., Biemann C., Panchenko A.* Negative Sampling Improves Hypernymy Extraction Based on Projection Learning // Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017): Volume 2, Short Papers, April 3–7, 2017, Valencia, Spain. Stroudsburg, PA, USA: Association for Computational Linguistics, 2017. P. 543–550.
7. *Ustalov D.* Russian Thesauri as Linked Open Data // Computational Linguistics and Intellectual Technologies: Papers from the Annual conference "Dialogue". Volume 1 of 2. Main conference program, May 27–30, 2015, Moscow, Russia. Moscow, Russia: RGGU, 2015. P. 616–625.

Статьи в других изданиях

8. *Ustalov D., Panchenko A., Biemann C.* Watset: Automatic Induction of Synsets from a Graph of Synonyms // Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017) (Volume 1: Long Papers), July 30 – August 4, 2017, Vancouver, BC, Canada. Stroudsburg, PA, USA: Association for Computational Linguistics, 2017. P. 1579–1590.

Свидетельства о регистрации программ для ЭВМ

9. *Усталов Д.* Свидетельство Роспатента о государственной регистрации программы для ЭВМ «Программа подбора проекционной матрицы для векторных представлений слов» № 2017615703 от 22.05.2017.

Исследование выполнено при финансовой поддержке
РФФИ в рамках научного проекта № 16-37-00354 мол_а,
РГНФ в рамках научных проектов № 13-04-12020 и 16-04-12019,
и стипендии Президента Российской Федерации № СП-773.2015.5.

Подписано в печать ___.___._____. Заказ № _____
Формат 60×90/16. Усл. печ. л. 1. Тираж 100 экз.
Типография _____