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

05.13.17 — теоретические основы информатики

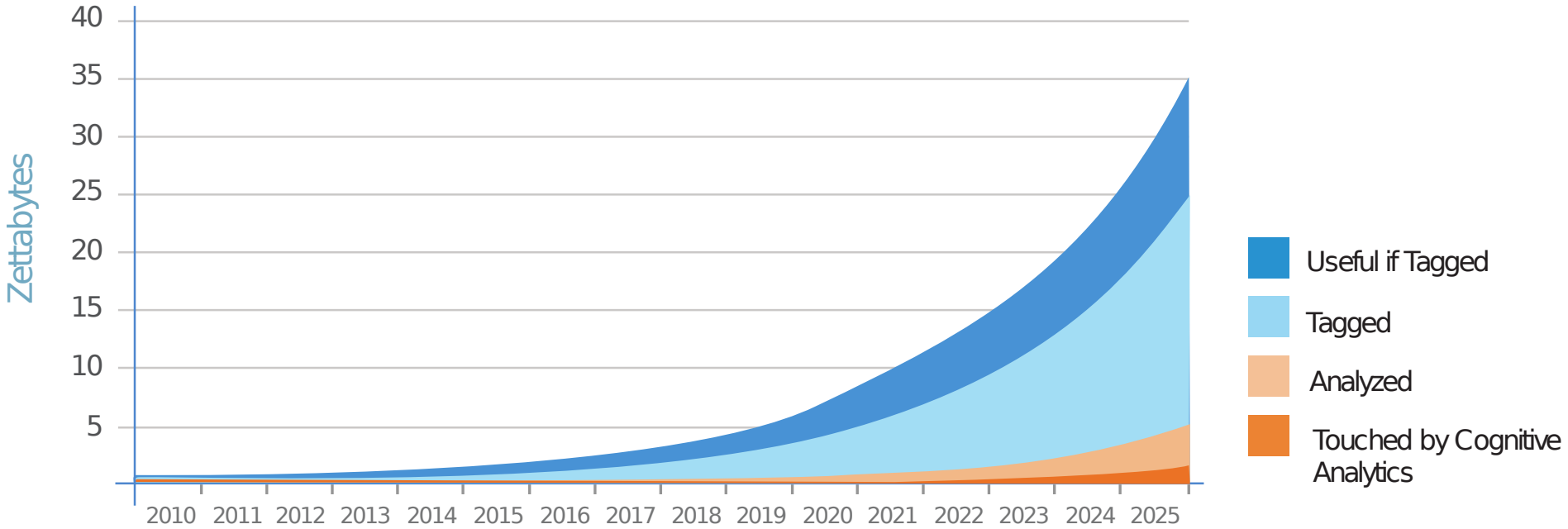
Диссертация на соискание ученой степени кандидата физико-математических наук

Усталов Дмитрий Алексеевич

ИММ УрО РАН, Екатеринбург

Научный руководитель:
Созыкин Андрей Владимирович,
к. т. н.

Проблема структурирования данных

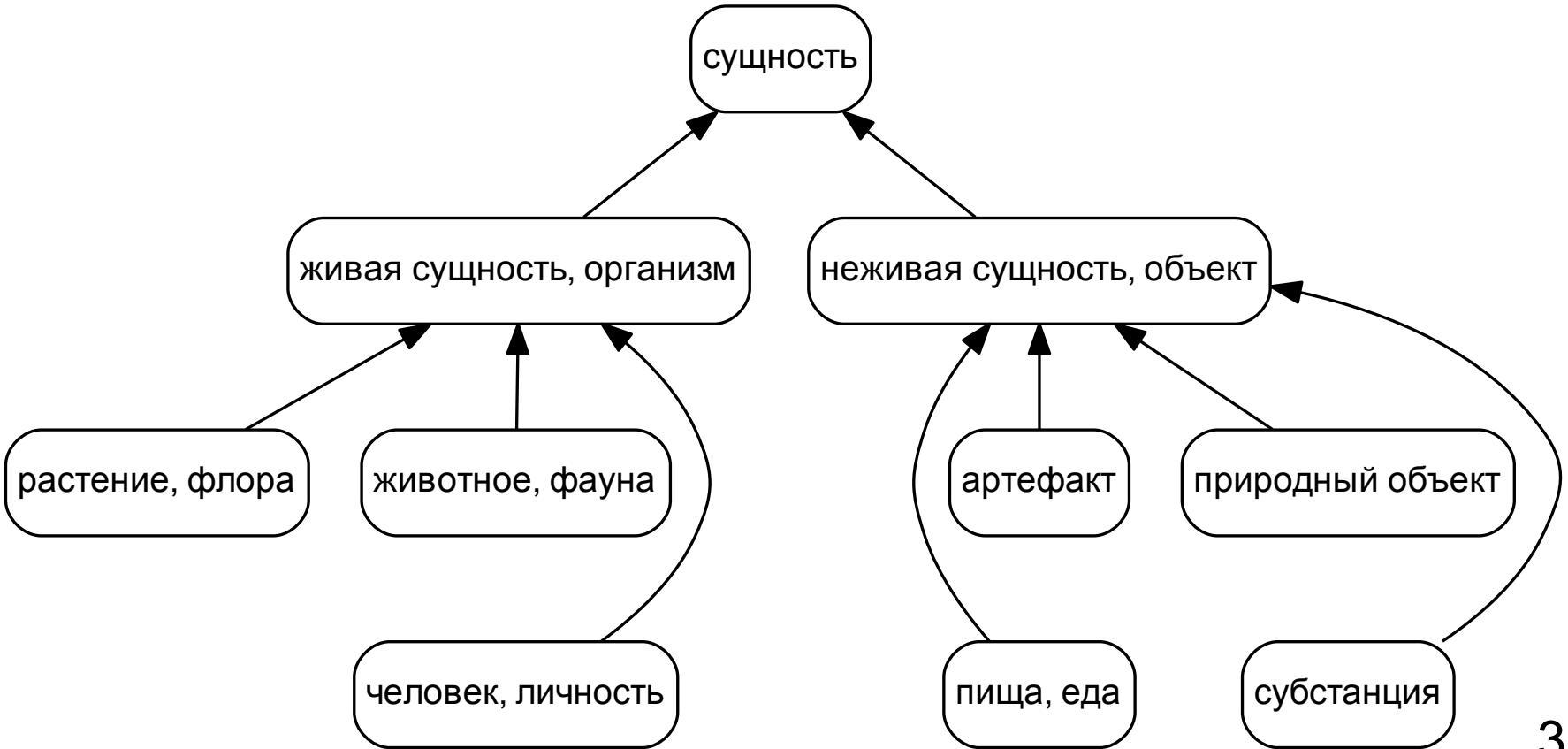


Source: IDC's Data Age 2025 study, sponsored by Seagate, April 2017

- К 2025 г. только $15\ \%$ всех данных в мире будут размеченными, причем только $20\ \%$ размеченных данных будут проанализированы
- Основную часть данных составляют неструктурированные данные: текст, изображения, аудио, видео
- Структурирование текстовых данных затрудняется **лексической многозначностью**

Развитие технологий обработки текста

- Применение методов машинного обучения для обнаружения закономерностей в данных
- Использование семантических сетей для разрешения лексической многозначности



Цель диссертационной работы

Цель работы: 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.

Основные задачи

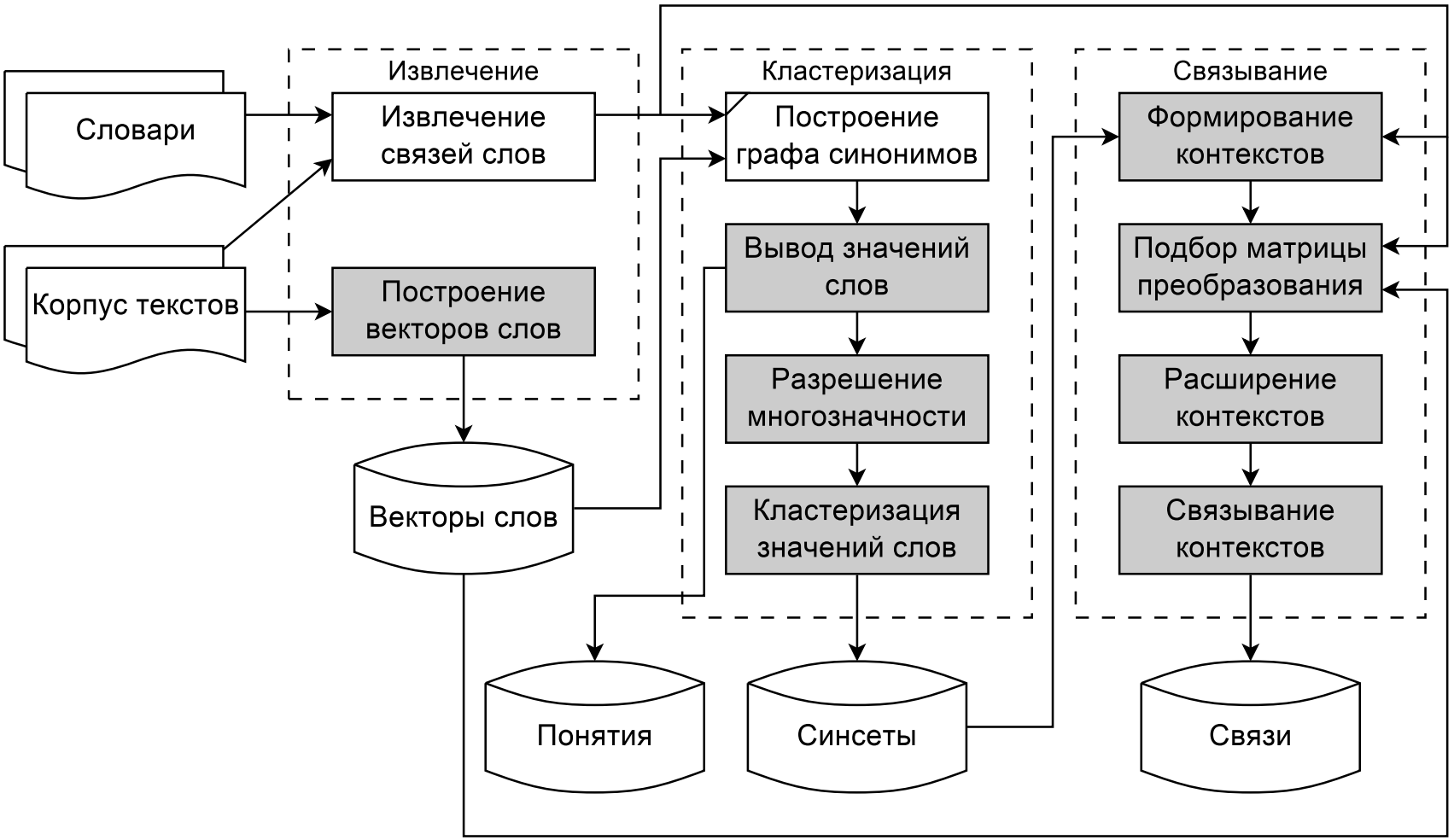
1. Разработать математическую модель представления лексических значений слов и связей между ними в виде семантической сети слов
2. Разработать метод и алгоритм построения синсетов на основе разрешения многозначности слов
3. Разработать метод и алгоритм построения и расширения однозначных семантических 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

Работы по теме диссертации

Понятия	*Dorow B., Widdows D.* Discovering Corpus-Specific Word Senses // Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics - Volume 2. EACL '03. ACL, 2003. P. 79–82.	Отсутствует группировка близких слов
	Hope D., Keller B. MaxMax: A Graph-Based Soft Clustering Algorithm Applied to Word Sense Induction // Computational Linguistics and Intelligent Text Processing: 14th International Conference, CICLing 2013, Proceedings, Part I. Springer, 2013. P. 368–381.	Зависимость от распределения весов ребер
	Palla G., Derenyi I., Farkas I., Vicsek T. Uncovering the overlapping community structure of complex networks in nature and society // *Nature*. 2005. Vol. 435. P. 814–818.	Недостижимое допущение о структуре клик
Связи	*Hearst M. A.* Automatic Acquisition of Hyponyms from Large Text Corpora // Proceedings of the 14th Conference on Computational Linguistics - Volume 2. COLING '92. ACL, 1992. P. 539–545.	Сильная разреженность и зашумленность
	Fu R., Guo J., Qin B. et al. Learning Semantic Hierarchies via Word Embeddings // Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). ACL, 2014. P. 1199–1209.	Отсутствие связей между значениями слов
	Крижановский А. А., Смирнов А. В. Подход к автоматизированному построению общецелевой лексической онтологии на основе данных Викисловаря // *Известия РАН. Теория и системы управления*. 2013. № 2. С. 53–63.	Отсутствие связей между значениями слов
Сеть	*Gonçalo Oliveira H., Gomes P.* ECO and Onto.PT: a flexible approach for creating a Portuguese wordnet automatically // *Language Resources and Evaluation*. 2014. Vol. 48, no. 2. P. 373–393.	Зависимость от внешнего ресурса

Предлагаемое решение

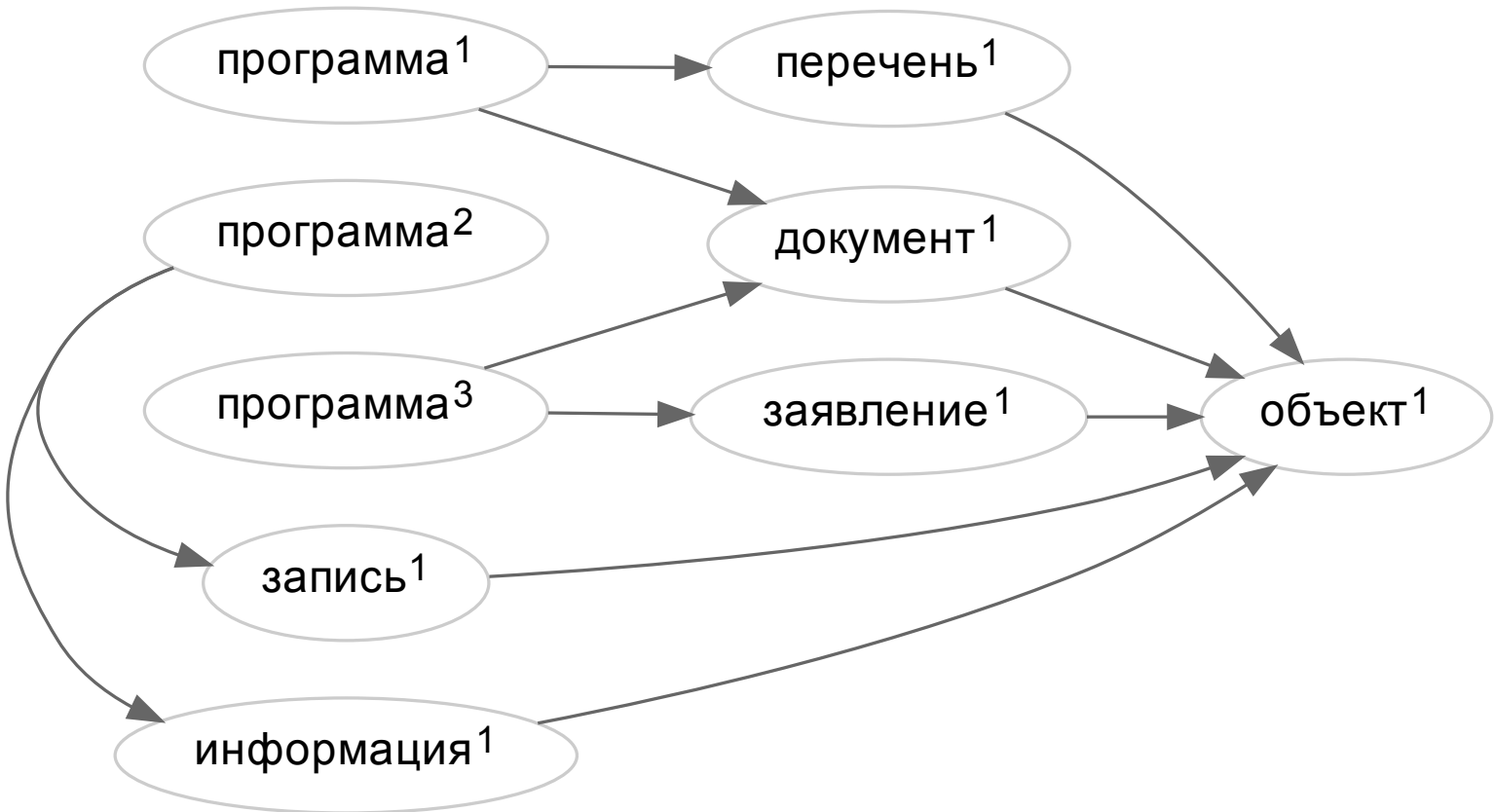
- Формировать связи между отдельными значениями слов
- Не использовать внешние базы знаний
- Использовать слабоструктурированные данные



Семантическая сеть слов

Определение

Семантическая сеть слов $\mathcal{N} = (\mathcal{V}, \mathcal{R})$ — это ориентированный граф, вершины которого — лексические значения слов $\mathcal{V}$, множество дуг $\mathcal{R} \subset \mathcal{V} \times \mathcal{V}$ которого порождается асимметричным отношением на множестве $\mathcal{V}$.



Словник и граф синонимов

Определение

Словник V — это множество всех лексических единиц заданного языка.

Пусть задана некоторая **мера семантической близости** слов $\text{sim}_{\text{word}} : (u, v) \to \mathbb{R}, \forall u \in V, v \in V$.

Определение

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 близости sim_{word}.

Построение синсетов

Определение

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

- Кластеризация графа $W = (V, E)$ затруднена тем, что словник V содержит однозначные и многозначные слова
- Пусть граф W можно преобразовать в такой граф $\mathcal{W} = (\mathcal{V}, \mathcal{E})$, что $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ порождается отношением синонимии на $\mathcal{V}$

Постановка задачи

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.

Метод построения синсетов Watset

Метод Watset

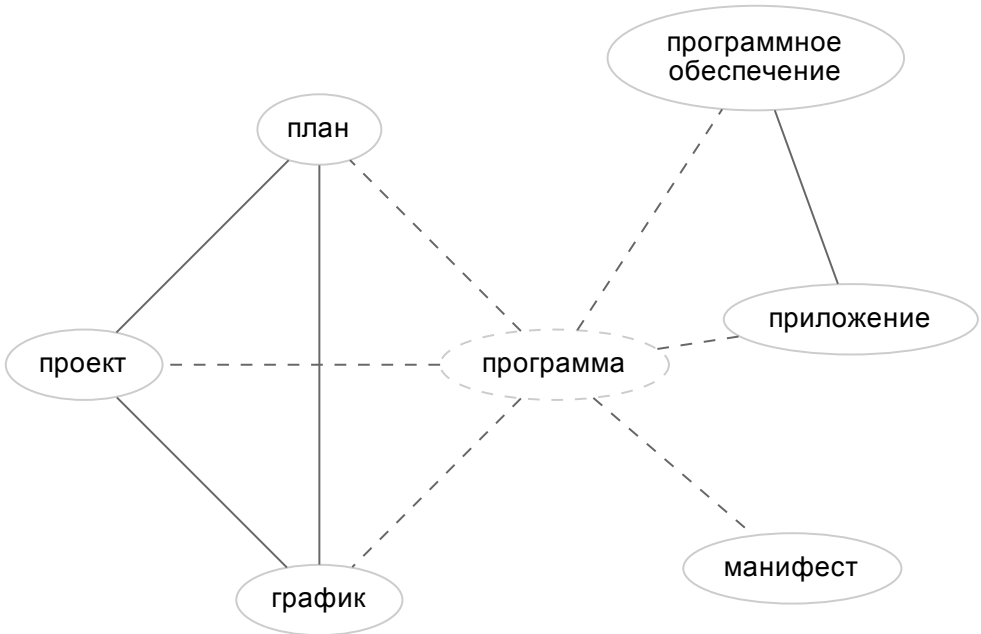
1. Вывести значения слов в V и построить множество $\mathcal{V}$.
2. Построить вспомогательный граф значений слов $\mathcal{W} = (\mathcal{V}, \mathcal{E})$.
3. Выполнить кластеризацию графа $\mathcal{W}$ при помощи какого-либо метода жесткой кластеризации графа.

Определение

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.

Метод Watset: вывод значений слова $u \in V$

- Извлечь и кластеризовать окрестность вершины $u \in V$ в графе $W = (V, E)$
- Записать кластеры C в **контексты** $\mathrm{ctx}(u^i) = C_i$, $1 \leq i \leq |C|$
- Записать значения $\mathrm{senses}(u) = \{u^i, 1 \leq i \leq |C|\}$



Значение	**Контекст**
программа^1	{*план*, *проект*, *график*}
программа^2	{*программное обеспечение*, *приложение*}
программа^3	{*манифест*}

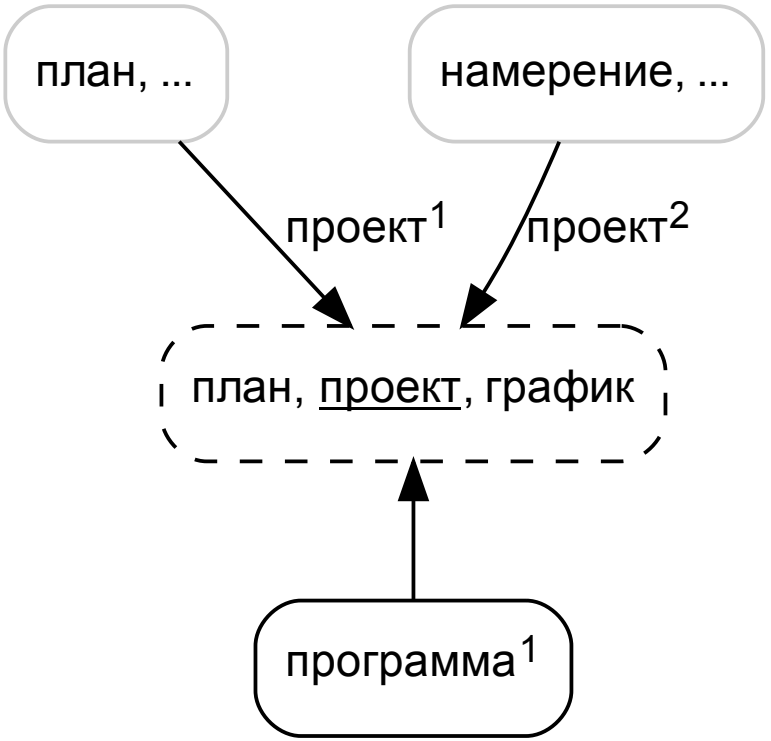
Формируется множество лексических значений слов $\mathcal{V} = \bigcup_{u \in V} \mathrm{senses}(u)$.

Метод Watset: разрешение многозначности

Пусть задана некоторая **мера близости** контекстов $\mathrm{sim}_{\mathrm{ctx}} : (\mathrm{ctx}(a), \mathrm{ctx}(b)) \to \mathbb{R}, \forall a \in \mathcal{V}, b \in \mathcal{V}$.

Нужно определить значения слов в контексте каждого значения слова $s \in \mathcal{V}$.

- Каждому элементу $u \in \mathrm{ctx}(s)$ ставится в соответствие $\hat{u} \in \mathcal{V}$:
  $$\hat{u} \in \underset{u' \in \mathrm{senses}(u)}{\arg\max}\ \mathrm{sim}_{\mathrm{ctx}}(\mathrm{ctx}(s), \mathrm{ctx}(u'))$$
- Однозначный контекст:
  $$\widehat{\mathrm{ctx}}(s) = \{\hat{u} : u \in \mathrm{ctx}(s)\}$$



Теорема

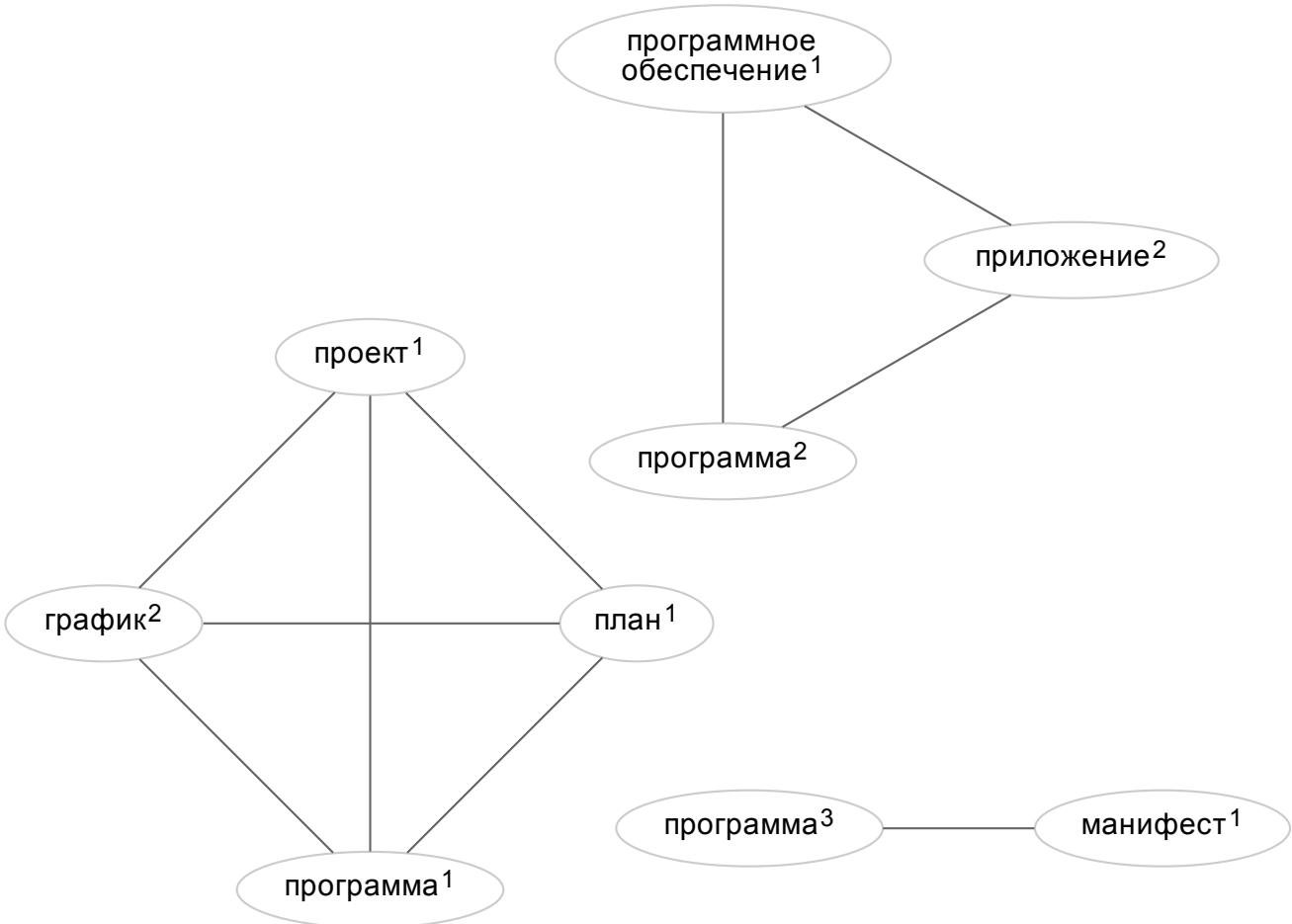
Пусть $\deg_{\max}$ *— максимальная степень вершины графа* $W = (V, E)$. *Тогда вычислительная сложность процедуры разрешения многозначности контекста всех значений слов составляет* $O(|V| \deg^4_{\max})$ *при* $\mathrm{sim}_{\mathrm{ctx}} = \cos$.

Метод Watset: построение графа значений слов

На основе контекстов со снятой многозначностью $\widehat{\mathrm{ctx}}(s) \subset \mathcal{V}$, $\forall s \in \mathcal{V}$ формируется множество ребер $\mathcal{E}$ графа значений слов $\mathcal{W}$:

$$\mathcal{E} = \{\{\hat{u}, \hat{v}\} \in \mathcal{V} \times \mathcal{V} : \hat{v} \in \widehat{\mathrm{ctx}}(\hat{u})\}.$$

- Производится кластеризация графа $\mathcal{W} = (\mathcal{V}, \mathcal{E})$
- Полученное множество кластеров $\mathcal{S}$ является искомым множеством синсетов



Построение связей

Определение

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 $h \in V$.

R	
w	h
программа	документ
программа	запись
план	документ
план	перечень
приложение	запись

Определить значения слов в парах $(w, h) \in R$ трудно.

Метод построения связей Watlink

Постановка задачи

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 $s \in S$.

Пусть $\text{words}(S) \subseteq V$ — множество слов, значения которых включены в синсет S.

Метод Watlink

1. Построить иерархический контекст для каждого синсета $S \in \mathcal{S}$.
2. Расширить иерархические контексты синсетов.
3. Разрешить многозначность в иерархических контекстах.

Допущение: $|\bigcap_{w \in \text{words}(S)} \{h \in V : (w, h) \in R\}| > 0, \forall S \in \mathcal{S}$.

Метод Watlink: иерархический контекст

Определение

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

- Каждому синсету $S \in \mathcal{S}$ ставится в соответствие $\mathrm{hctx}(S) = \{h \in V : (w, h) \in R, w \in \mathrm{words}(S), h \notin \mathrm{words}(S)\}$
- Важность слов в контекстах различается: выполняется взвешивание $\mathrm{tf\text{–}idf}(h, S, \mathcal{S}) = \mathrm{tf}(h, S) \times \mathrm{idf}(h, \mathcal{S})$

Синсет	**Иерархический контекст**
{*программа*^{1}, *план*^{1}, ...}	{*документ*, *перечень*, ...}
{*программа*^{2}, *приложение*^{2}, ...}	{*запись*, *информация*, ...}
{*программа*^{3}, *манифест*^{1}, ...}	{*документ*, *заявление*, ...}

На практике, доступность данных для построения отношения R низка.

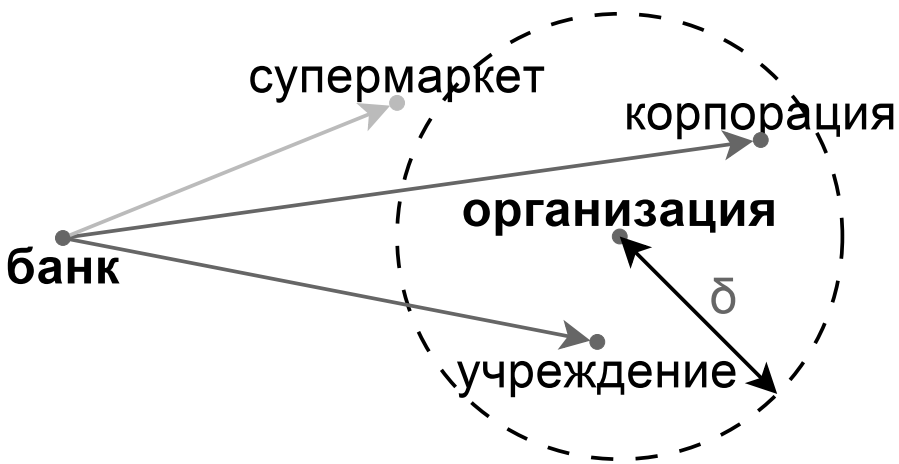
Метод Watlink: расширение контекста для $S \in \mathcal{S}$

Пусть задано $\vec{u} : u \to \mathbb{R}^d, d \in \mathbb{N}, |V| \gg d, \forall u \in V$.

Пусть $\mathrm{NN}_n(\vec{u}) \in V$ — операция поиска $n \in \mathbb{N}$ ближайших соседей слова $u \in V$ в $\mathbb{R}^d$.

Пусть Φ^* — такая матрица, что $\Phi^* \vec{w} = \vec{h}, \forall (w, h) \in R$.

- Построение множества кандидатов в контекст:
  $M_S = \bigcup_{h \in \mathrm{hctx}(S)} \mathrm{NN}_n(\vec{h}) \setminus \mathrm{hctx}(S)$
- Проверка $h \in M_S$: если задано $\delta \in \mathbb{R}^+$ и выполняется условие
  $\exists w \in \mathrm{words}(S) : \|\Phi^* \vec{w} - \vec{h}\| < \delta$,
  то $\mathrm{hctx}(S) = \mathrm{hctx}(S) \cup \{h\}$



Метод Watlink: подбор элементов матрицы Φ^*

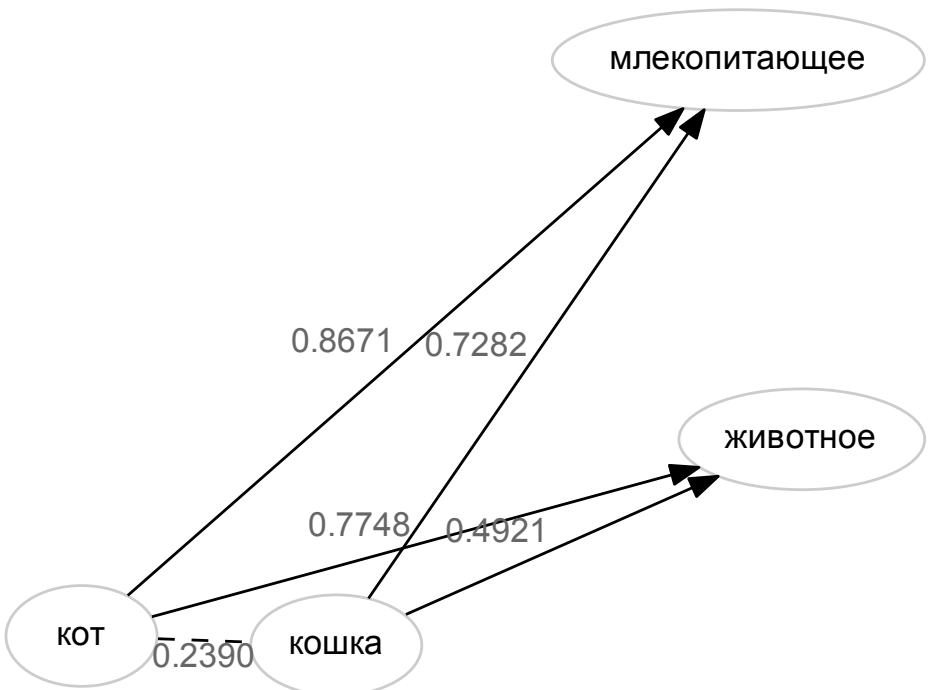
Элементы матрицы Φ^* подбираются на основе упорядоченных пар слов $(w, h) \in R$ методом наименьших квадратов с использованием стабилизатора H:

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

Стабилизатор H учитывает асимметричность отношения R:

$$H = \sum_{(\vec{w},\vec{h})\in R} ((\Phi^2\vec{w})^T\vec{w})^2.$$

Результат *повторного* умножения Φ^* на $\vec{w}$ не должен быть близок к $\vec{w}$.



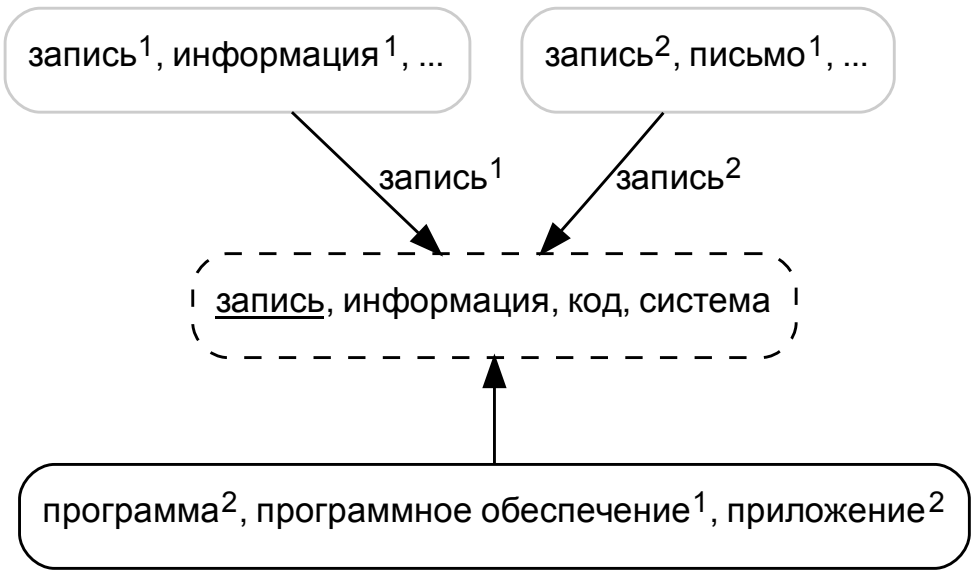
Метод Watlink: разрешение многозначности

Пусть задана некоторая **мера близости** иерархического контекста и слов синсета $\mathrm{sim}_{\mathrm{hctx}} : (\mathrm{hctx}(A), \mathrm{words}(B)) \to \mathbb{R}, \forall A \in \mathcal{S}, B \in \mathcal{S}$.

Нужно определить значения слов в иерархическом контексте каждого синсета $S \in \mathcal{S}$.

- Каждому элементу $h \in \mathrm{hctx}(S)$ ставится в соответствие $\hat{h} \in \mathcal{V}$:
  $\hat{h} \in \arg\max_{h' \in \mathrm{senses}(h): S' \in \mathcal{S}, h' \in S', S \neq S'} \mathrm{sim}_{\mathrm{hctx}}(\mathrm{hctx}(S), \mathrm{words}(S'))$
- Однозначный контекст:
  $\widehat{\mathrm{hctx}}(S) = \{\hat{h} : h \in \mathrm{hctx}(S)\}$

Выполненные построения позволяют сформировать **семантическую сеть слов** $\mathcal{N} = (\mathcal{V}, \mathcal{R})$, где
$\mathcal{V} = \bigcup_{u \in V} \mathrm{senses}(u)$,
$\mathcal{R} = \bigcup_{S \in \mathcal{S}} S \times \widehat{\mathrm{hctx}}(S)$.



Разработка комплекса программ

Разработан комплекс программ SWN, реализующий предложенные модели, методы и алгоритмы.

- Языки программирования: Python, AWK и Bash
- Использованные библиотеки: scikit-learn, Gensim, TensorFlow, NetworkX, Raptor
- Исходные тексты доступны в сети Интернет:

  github.com/dustalov/watset

  github.com/dustalov/projlearn

  github.com/dustalov/watlink

Оценка эффективности методов

Эксперименты

- Оценка эффективности метода построения синсетов
- Оценка эффективности метода подбора матрицы ЛП
- Оценка эффективности метода построения связей

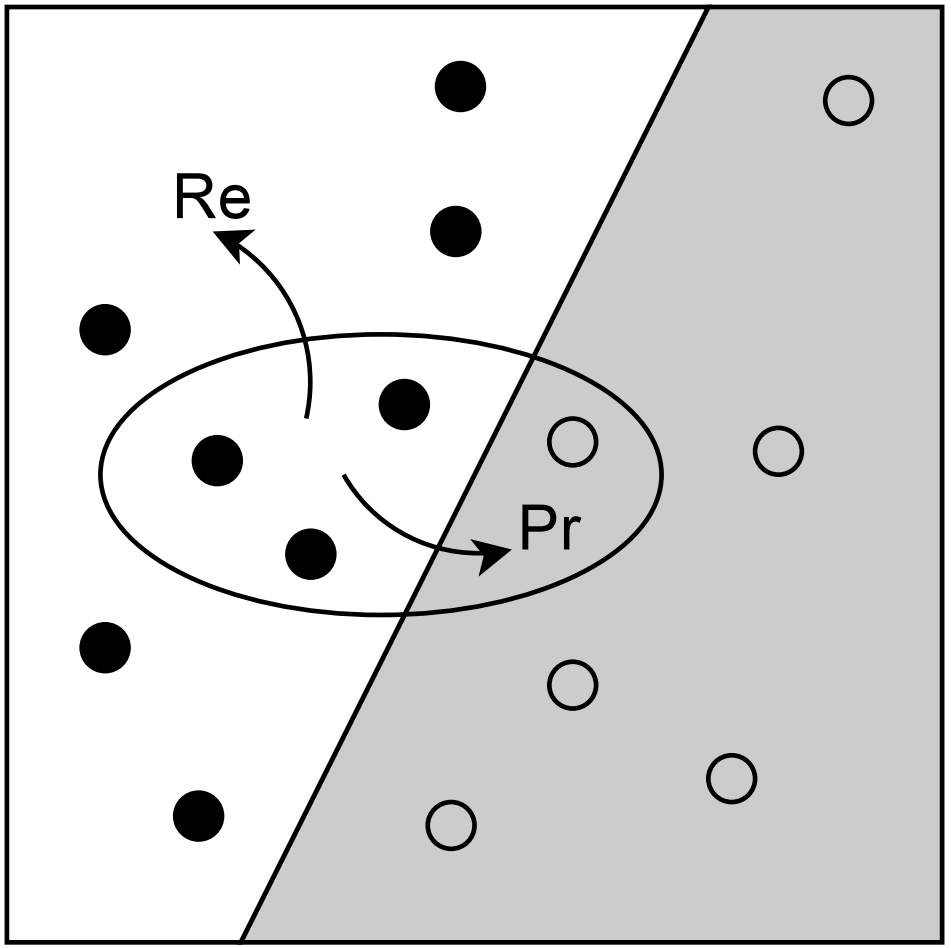
Виртуальная машина NC24 в облачной среде Microsoft Azure	
Тип центрального процессора	Intel Xeon E5-2690 v3
Количество доступных ядер	24 ядра
Объем оперативной памяти	224 ГБ
Тип графического процессора	NVIDIA Tesla K80
Объем видеопамяти	12 ГБ
Операционная система	CentOS 7.0 (64 бит, Linux)

Подход к оценке эффективности: золотой стандарт

- Информационно-поисковые критерии качества:

$$\mathrm{Pr} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}},$$
$$\mathrm{Re} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}},$$
$$\mathrm{F}_1 = 2\frac{\mathrm{Pr} \cdot \mathrm{Re}}{\mathrm{Pr} + \mathrm{Re}},$$

  где TP — верные положительные ответы, FP — ложные положительные ответы, FN — ложные отрицательные ответы



- Критерий качества hit@k

Оценка метода Watset: описание эксперимента

Оценивается качество построения синсетов путем сопоставления связанных пар слов с золотым стандартом.

- **Меры качества:** точность, полнота и F_1-мера
- **Золотой стандарт:** RuWordNet, Yet Another RussNet
- **Входные данные:** граф синонимов ($83\,092$ вершин и $211\,986$ ребер) на основе наборов данных:

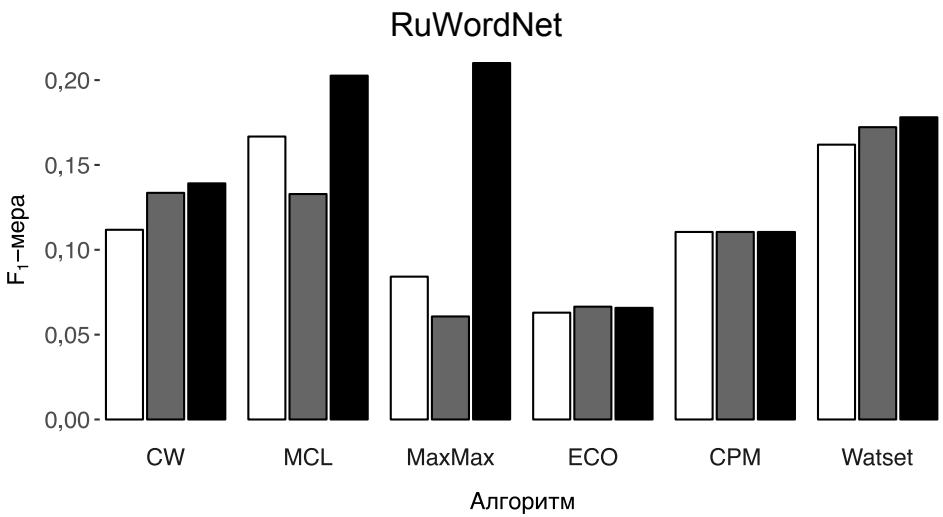
  Словарь синонимов Н. А. Абрамова

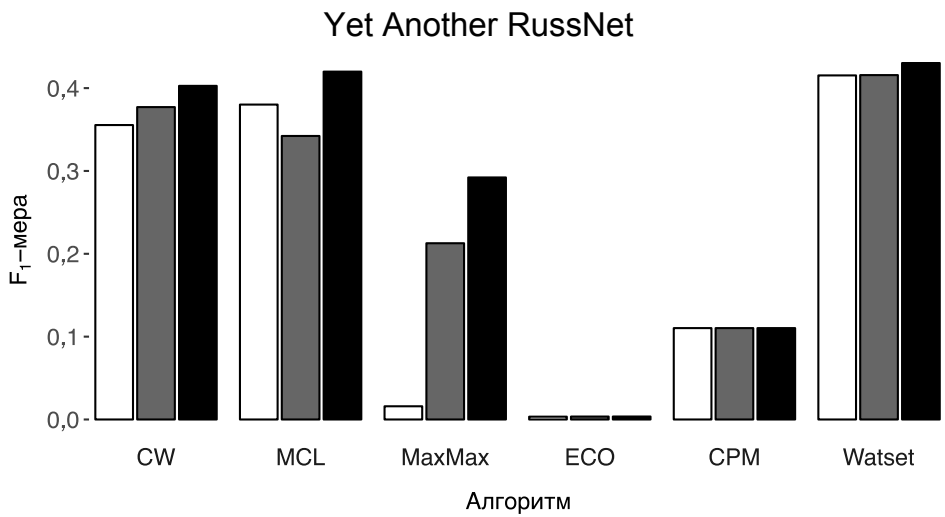
  Русский Викисловарь

  Универсальный словарь концептов (UNLDC)

Оценка метода Watset: результаты эксперимента

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
$CPM_{k=3}$	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





Обозначения: единичные веса ребер; кол-во словарей синонимов; семантическая близость

Подбор матрицы ЛП: результаты эксперимента

Оценивается качество преобразования вектора нижестоящего слова в вектор вышестоящего слова при помощи матрицы ЛП.

- **Мера качества:** hit@k
- **Входные данные:**
  обучающая выборка: $25\,067$ пар слов (Викисл. + Шаблоны)
  проверочная выборка: $8\,192$ пар слов (Викисловарь)
  тестовая выборка: $8\,310$ пар слов (Викисловарь)

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**

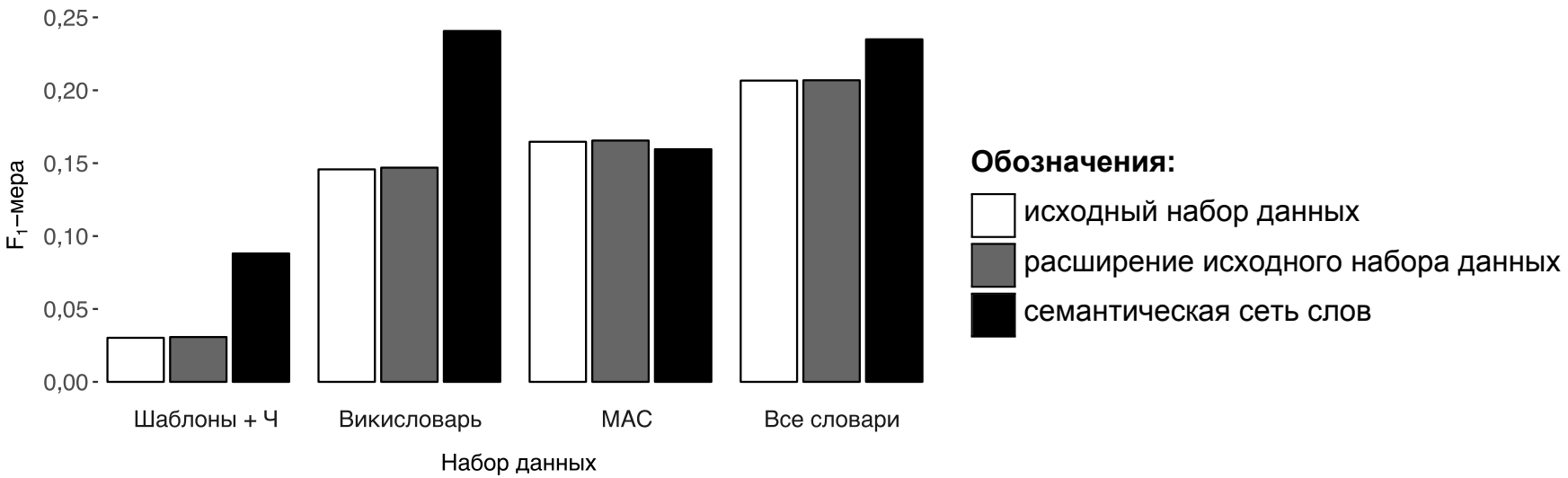
Оценка метода Watlink: описание эксперимента

Оценивается качество построения иерархических связей между значениями слов.

- **Меры качества:** точность, полнота и F_1-мера
- **Золотой стандарт:** RuWordNet
- **Входные данные:**
  «Шаблоны» — шаблоны Херст на материалах корпуса художественных текстов
  «Шаблоны + Ч» — «Шаблоны» с удалением низкочастотных пар слов
  «Викисловарь» — родо-видовые пары слов из Викисловаря
  «МАС» — шаблоны Херст на толкованиях Малого академического словаря
  «Все словари» — объединение трех ресурсов: «Шаблоны + Ч», «Викисловарь» и «МАС»

Оценка метода Watlink: результаты эксперимента

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**



Основные результаты, выносимые на защиту

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