Методы повышения эффективности процесса коллективного построения лексических ресурсов

Д. А. Усталов
Научный руководитель: **А. В. Созыкин**

Языковые ресурсы

Ресурсы: словари, тезаурусы, корпусы текстов, и т. д.

Определение

Тезаурус *— словарь, в котором слова и словосочетания с близкими значениями сгруппированы в единицы, называемые понятиями, и в котором явно указываются семантические отношения между этими понятиями.*

Применение электронных тезаурусов:

- снятие семантической неоднозначности;
- расширение поисковых запросов;
- анализ вопросов в системах общения;
- и др.

Тезаурусы русского языка

Ресурс	Год	Понятий	Отношений	Лицензия
RussNet	1999	30 тыс.	45 тыс.	—
РуТез	2002	55 тыс.	210 тыс.	CC BY-NC-SA*
WordNet.Ru	2003	31 тыс.	—	WordNet
Russian Wordnet	2004	—	—	—
Викисловарь	2004	190 тыс.*	70 тыс.*	CC BY-SA
UNL	2012	62 тыс.	90 тыс.	CC BY-SA
BabelNet*	2012	2,5 млн.	380 млн.	CC BY-NC-SA
YARN	2013	69 тыс.	30 тыс.*	CC BY-SA

Проблемы

- Сложность и длительность процесса создания тезауруса.
- Высокие требования к квалификации лексикографов.
- Доступность и лицензирование существующих ресурсов.

Предмет, цели и задачи исследования

Предмет исследования

Процесс построения лексических ресурсов.

Цель исследования

Разработать эффективные методы построения лексических ресурсов при помощи краудсорсинга.

Задачи исследования

1. Разработка методики построения тезауруса при помощи краудсорсинга.
2. Разработка вычислительной модели для выполнения процедур разметки.
3. Разработка комплекса программ.

Краудсорсинг

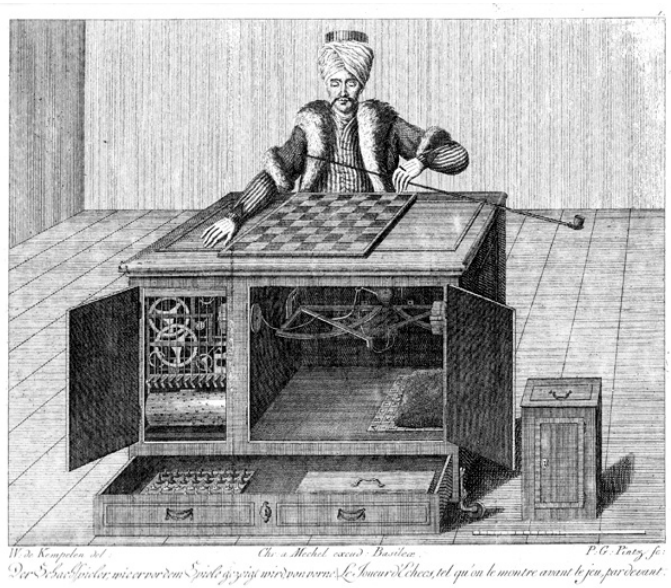
Определение

Краудсорсинг *— коллективный процесс решения задачи, поставленной заказчиком перед толпой участников на специализированной человеко-машинной платформе.*

Основное внимание в данной работе посвящено краудсорсингу *микрозадачами*.

Обзор работ:

- методики решения задач;
- методы обеспечения качества;
- программное обеспечение;
- подходы к оценке качества.



Методики решения задач

Исследователи разрабатывают методики для решения задач при помощи краудсорсинга.

- **ESP Game** (2006) — игрофицированная разметка изображений.
- **Find-Fix-Verify** (2010) — вычитка, перефразирование и улучшение документов Microsoft Word.
- **CrowdForge** (2011) — адаптация модели вычислений MapReduce.
- **CrowdDB** (2011) — оператор `CROWD` для размещения задач при выполнении SQL-запроса.
- **CrowdWeaver** (2012) — потоковая модель вычислений для краудсорсинга.
- **TWSI** (2013) — построение языковых ресурсов на основе явления лексической замещаемости.

Методы обеспечения качества

Оплата труда не гарантирует высокого качества результата.
Создаются методы вероятностного вывода.

- Метод **Давина-Скина** (1979) — построение матриц ошибок при помощи EM-алгоритма.
- **GLAD** (2009) — вывод ответов, сложности заданий и квалификации участников на основе EM-алгоритма.
- Алгоритм **Каргера-Оха-Шаха** (2011) — оптимальное по порядку назначение заданий и вывод ответов.
- **ZenCrowd** (2012) — вывод ответов и квалификации участников на основе фактор-графов.
- ***i*Crowd** (2015) — онлайн-алгоритм назначения заданий, оценки участников и вывода ответов.

Программное обеспечение

Создаётся программное обеспечение для упрощения запуска микрозадач и обработки результатов работы.

- **TurKit** (2009) — инструменты для упрощения размещения и выполнения заданий на MTurk.
- **SQUARE** (2013) — средства оценки качества методов обеспечения качества.
- **WebAnno** (2013) — среда для разметки текстов при помощи краудсорсинга.
- **ActiveCrowdToolkit** (2015) — средства оценки качества методов обеспечения качества.
- **CEKA** (2015) — средства анализа результатов выполнения микрозадач.
- **psiTurk** (2015) — сервис организации воспроизводимого процесса разметки.

Подходы к оценке качества

Подходы

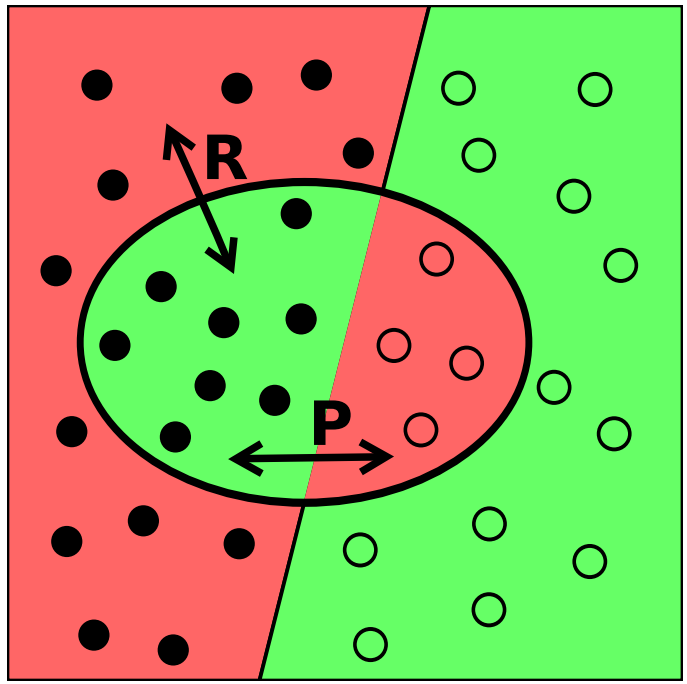
Золотой стандарт, экспертная оценка, разделяемая задача.

Ответы:

- верные положительные (*TP*),
- верные отрицательные (*TN*),
- ложные положительные (*FP*),
- ложные отрицательные (*FN*).

Метрики бинарной классификации:

- Точность $P = \frac{TP}{TP+FP}$.
- Полнота $R = \frac{TP}{TP+FN}$.
- F_1-мера $F_1 = 2\frac{P \cdot R}{P+R}$.



Информационно-поисковый подход.

Методика построения тезауруса предметной области I

Постановка задачи

- **Дано:** словарь предметной области $\mathbb{D}$, множество понятий $\mathbb{S}$, множество родовидовых пар слов $\mathbb{R}$ неизвестного качества.
- **Необходимо:** провести очистку данных и построить тезаурус предметной области.

Методика построения тезауруса предметной области

1. Извлечь из $\mathbb{S}$ все понятия, содержащие слова из $\mathbb{D}$.
2. Извлечь из $\mathbb{R}$ все представленные пары слова.
3. Построить родовидовые отношения между $\mathbb{S}$ и $\mathbb{R}$.
4. Выполнить уточнение лексикализации понятий в $\mathbb{S}$.
5. Объединить дубликаты-когипонимы в $\mathbb{S}$.

Методика построения тезауруса предметной области II

Проблема 1

Синсеты получены из источника неизвестного качества и содержат ошибки.

- Недостающие слова: $\{\textit{мундир}, \textit{униформа}\} \leftarrow \{\textbf{\textit{форма}}\}$.
- Посторонние слова: $\{\textit{знать}, \textit{понимать}, \textbf{\textit{аристократия}}\}$.

Проблема 2

Формирование родовидовых отношений между синсетами на основе пар слов приводит к неоднозначности.

- *Отношение:* $(\textit{ткань}, \textit{джинса})$.
- *Синсеты:* $\{\textit{джинса}, \textit{реклама}\}$, $\{\textit{джинсовая ткань}, \textit{джинса}\}$.

Коллективные потоковые вычисления

Краудсорсинг микрозадачами предполагает обработку некоторого набора исходных данных в один или несколько связанных друг с другом этапов.

Определение

Коллективные потоковые вычисления *— потоковая вычислительная модель, объединяющая человеко-машинные этапы обработки реляционных данных.*

- Данные выражаются расширенной реляционной моделью с операциями вкладывания и выкладывания.
- Процедура решения задачи записывания в виде *схемы коллективных вычислений*.

Схема коллективных вычислений I

Определение

Схема коллективных вычислений *— слабо связный ориентированный ациклический граф W со множеством рёбер E, множество вершин которого образуется объединением множества этапов разметки S, множества этапов синхронизации Y и множества источников данных D.*

$$W = (S \cup Y \cup D, E) \qquad (1)$$

Условные обозначения:

- множество элементов схемы: $V = S \cup Y \cup D$;
- реляционное отношение $v \in V$: $\big(H(v), B(v)\big)$;
- первичный ключ элемента v: PK$(v) \subseteq H(v)$;
- множество входящих вершин в вершину v: *In*$(v) \subset V$.

Схема коллективных вычислений II

Определение

Этап разметки $s \in S$ *— это реляционное отношение, тело которого получено путём преобразования толпой участников кортежей единственного входящего отношения:* $In(s) \subset V \wedge |In(s)| = 1$.

Определение

Этап синхронизации $y \in Y$ *— это реляционное отношение, тело которого получено путём автоматической обработки двух и более входящих отношений:* $In(y) \subset V \wedge |In(y)| > 1$.

Определение

Источник данных $d \in D$ *— это реляционное отношение, тело которого получено заранее и не зависит от других элементов схемы коллективных вычислений:* $In(d) = \varnothing$.

Выполнение схем коллективных вычислений I

Определение

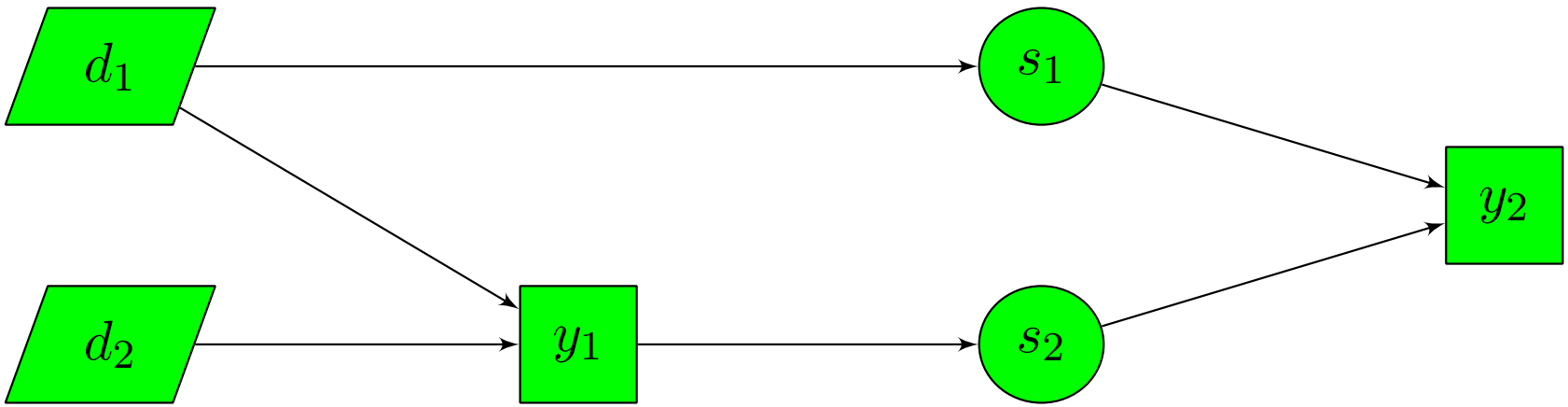
Согласованная схема коллективных вычислений *— схема, каждый кортеж каждого элемента которой однозначно идентифицирует породившие его кортежи.*

$$\forall v \in V \Big(\forall v' \in In(v) \big(H(v) \cap H(v') \supseteq \mathsf{PK}(v') \big) \Big) \qquad (2)$$

Алгоритм 1 Синхронный алгоритм выполнения, $|D| > 0$

1: **for all** $v \in V$ **do**
2: $\quad M_v \leftarrow (v \in D)$
3: **end for**
4: **parallel for all** $v \in V \setminus D$ **do**
5: $\quad \mathtt{Wait}(\forall v' \in \mathit{In}(v)(M_{v'} = \textbf{true}))$
6: $\quad B(v) \leftarrow \mathsf{Run}(v)$
7: $\quad M_v \leftarrow \textbf{true}$
8: **end for**

Выполнение схем коллективных вычислений II



$$M = [1, 1, 1, 1, 1, 1]$$

Уточнение лексикализации понятий I

Определение

Синсет *(синонимический ряд) — множество квазисинонимов, выражающих понятие.*

Пример

$\{$*автомобиль*, *машина*, *автó*, . . . $\}$

Постановка задачи

- **Дано:** множество синсетов $\mathbb{S}$ и множество слов-кандидатов для включения в них $\mathbb{W}$.
- **Необходимо:** добавить в синсеты из $\mathbb{S}$ недостающие слова из $\mathbb{W}$ и удалить из их посторонние слова.

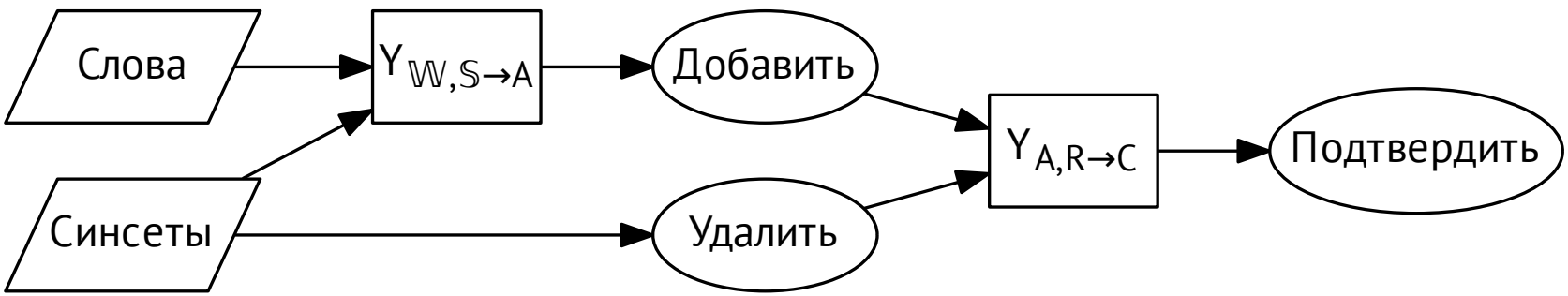
Процедура «добавить–удалить–подтвердить» I

Процедура *ARC*: «добавить–удалить–подтвердить»

Добавить: участник выбирает слова-кандидаты на включение в синсет.

Удалить: участник выбирает посторонние слова для удаления из синсета.

Подтвердить: участник выбирает между оригинальным синсетом и модифицированным.



Процедура «добавить–удалить–подтвердить» II

Элемент	Определение
$\mathbb{W}$	$H(\mathbb{W}) = \{\mathbb{S}.\mathit{id}, (\mathit{words}, \texttt{TEXT[]})\}$
$\mathbb{S}$	$H(\mathbb{S}) = \{(\underline{\mathit{id}}, \texttt{INT}), (\mathit{words}, \texttt{TEXT[]})\}$
$Y_{\mathbb{W},\mathbb{S}\to A}$	$\pi_{\mathbb{S}.\mathit{id},\mathit{words}=\mathbb{S}.\mathit{words},\ \mathit{candidates}=\mathbb{W}.\mathit{words}} (\mathbb{W} \bowtie \mathbb{S})$
A	$H(A) = \{(\underline{\mathit{id}}, \texttt{INT}), \mathbb{S}.\mathit{id}, (\mathit{added}, \texttt{TEXT[]})\}$
R	$H(R) = \{(\underline{\mathit{id}}, \texttt{INT}), \mathbb{S}.\mathit{id}, (\mathit{removed}, \texttt{TEXT[]})\}$
$Y_{A,R\to C}$	$\sigma_{\mathit{words}\neq\mathit{words}'} (\pi_{\mathbb{S}.\mathit{id},A.\mathit{id},R.\mathit{id},\mathit{words}=\mathbb{S}.\mathit{words},\ \mathit{words}'=\mathbb{S}.\mathit{words}\cup A.\mathit{added}\setminus R.\mathit{removed}} (\mathbb{S} \bowtie A \bowtie R))$
C	$H(C) = \{\mathbb{S}.\mathit{id}, A.\mathit{id}, R.\mathit{id}, Y_{A,R\to C}.\mathit{words}', (\mathit{b}, \texttt{BOOL})\}$

Построение родовидовых отношений

Определение

Родовидовое *(гипо-гиперонимическое)* ***отношение*** *— семантическое отношение между парой понятий, при котором одно понятие является разновидностью другого.*

Пример

$\{\textit{автомобиль}, \dots\} \overset{is-a}{\rightarrow} \{\textit{транспортное средство}, \dots\}$

Постановка задачи

- **Дано:** множество синсетов $\mathbb{S}$ и множество родовидовых пар слов $\mathbb{R}$.
- **Необходимо:** построить отношения между синсетами в $\mathbb{S}$ на основе $\mathbb{R}$.

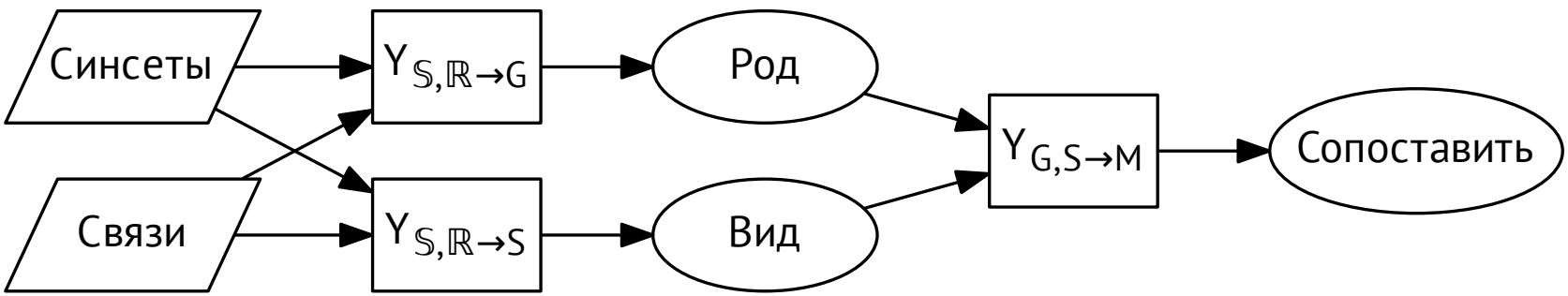
Процедура «род–вид–сопоставить» I

Процедура *GSM*: «род–вид–сопоставить>

Род: участник устанавливает синсет-род для пары *слов*.

Вид: участник устанавливает синсет-вид для пары *слов*.

Сопоставить: участник подтверждает осмысленность связи пары *синсетов*.



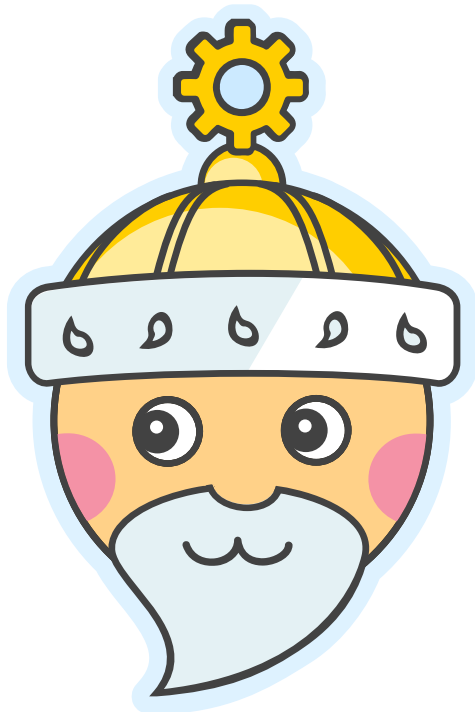
Процедура «род–вид–сопоставить» II

Элемент	Определение
$\mathbb{S}$	$H(\mathbb{S}) = \{(\underline{id}, \texttt{INT}), (words, \texttt{TEXT[]})\}$
$\mathbb{R}$	$H(\mathbb{R}) = \{(\underline{hypernym}, \texttt{TEXT}), (\underline{hyponym}, \texttt{TEXT})\}$
$Y_{\mathbb{S},\mathbb{R}\to G}$	$\pi_{\substack{id,words,\\ hypernym,hyponym}}\left(\mathbb{R} \bowtie \pi_{\substack{id,words,\\ hypernym=\mu(words)}}(\mathbb{S})\right)$
$Y_{\mathbb{S},\mathbb{R}\to S}$	$\pi_{\substack{id,words,\\ hypernym,hyponym}}\left(\mathbb{R} \bowtie \pi_{\substack{id,words,\\ hyponym=\mu(words)}}(\mathbb{S})\right)$
G	$H(G) = \{(\underline{id}, \texttt{INT}), \mathbb{S}.id, \mathbb{R}.hyponym, (b, \texttt{BOOL})\}$
S	$H(S) = \{(\underline{id}, \texttt{INT}), \mathbb{S}.id, \mathbb{R}.hypernym, (b, \texttt{BOOL})\}$
$Y_{G,S\to M}$	$\sigma_{G.b\wedge S.b=1}\left(\pi_{\substack{G.\mathbb{S}.id=s_1, S.\mathbb{S}.id=s_2,\\ G.id,S.id,G.b,S.b}}(G \bowtie S)\right)$
M	$H(M) = \{G.id, S.id, Y_{G,S\to M}.s_1, Y_{G,S\to M}.s_2, (b, \texttt{BOOL})\}$

Сервис управления процессом краудсорсинга

Сервис управления процессом краудсорсинга реализован в виде веб-сервиса на основе архитектуры REST.

- **Платформа:** Java 8.
- **Программный каркас:** Dropwizard (JAX-RS + Java EE).
- **Хранилище данных:** PostgreSQL.
- **Среда развёртывания:** Docker.



http://mtsar.nlpub.org/

https://github.com/mtsar/mtsar/

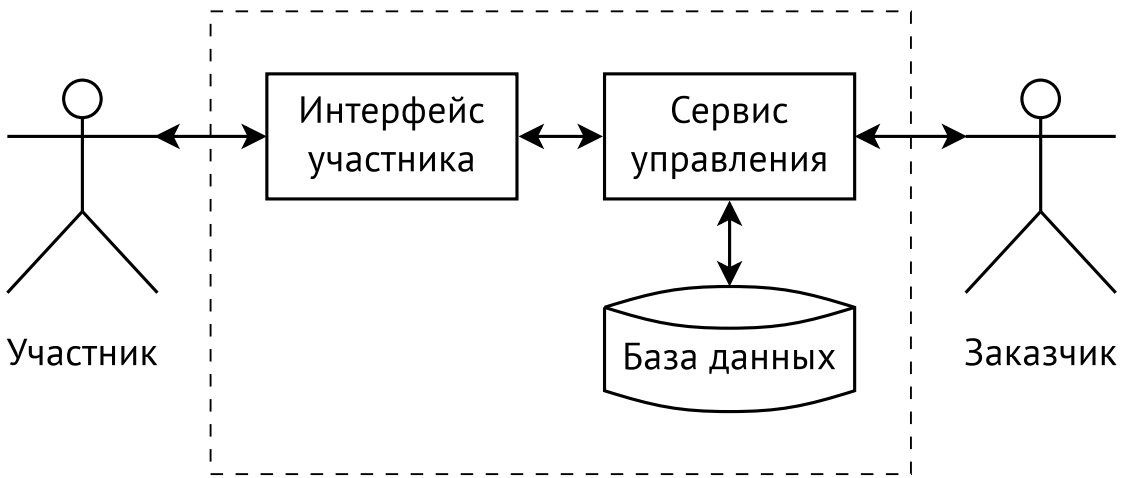
Функциональность комплекса программ

Интерфейс клиента:

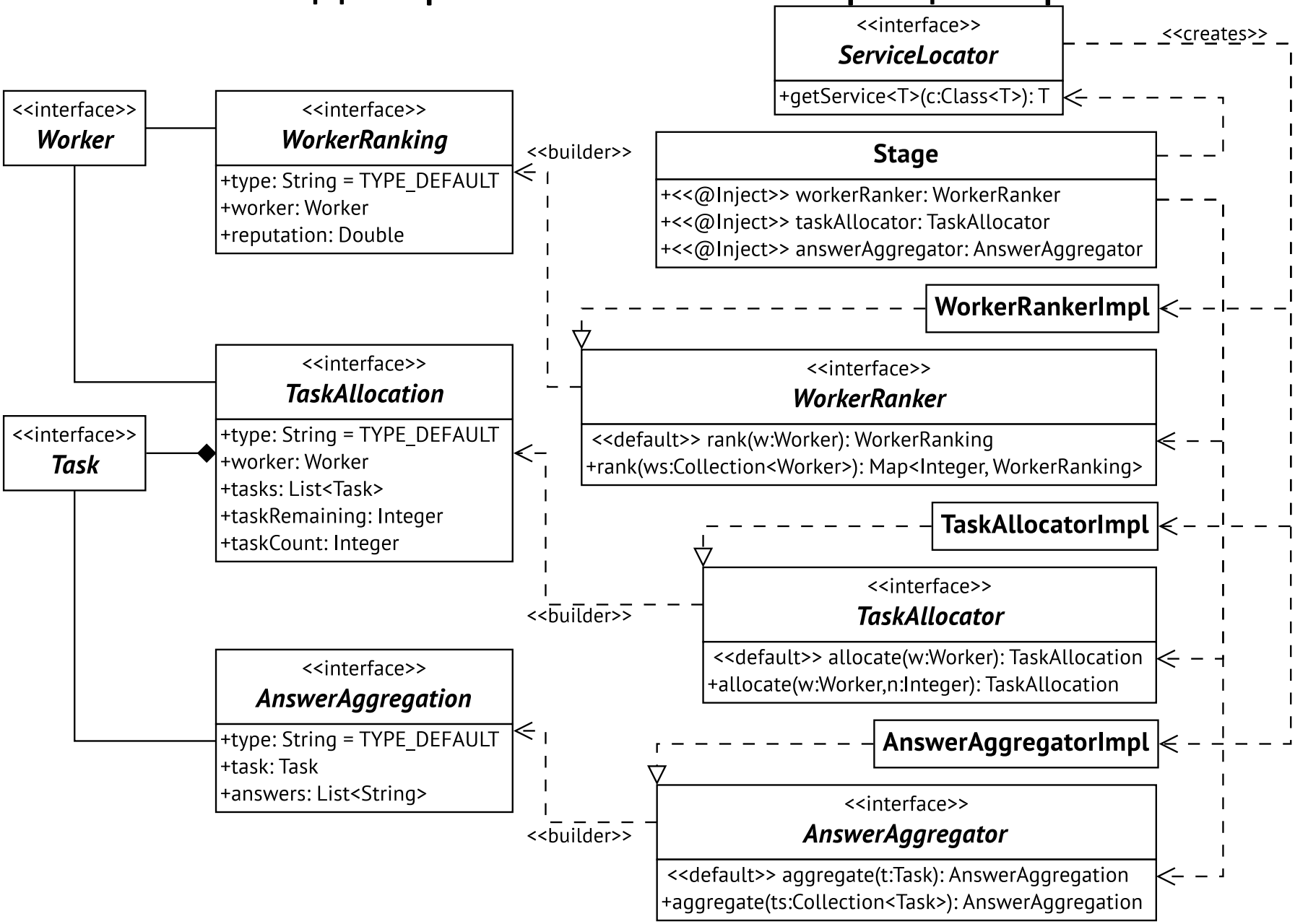
- запрос регистрации участника;
- запрос назначения задания;
- запрос приёма ответа.

Сервис управления:

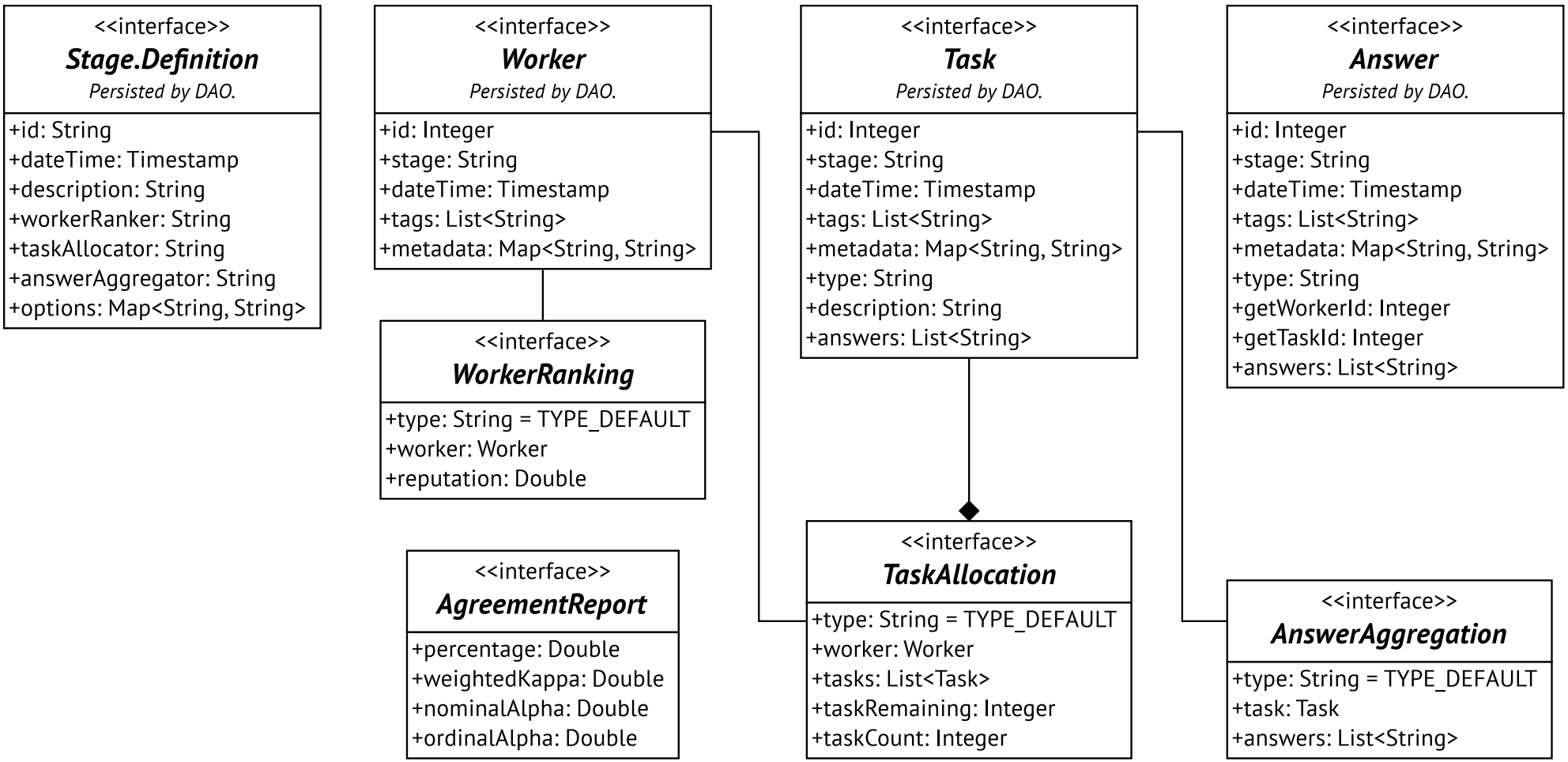
- назначение заданий;
- приём и агрегация ответов;
- ввод и вывод данных.



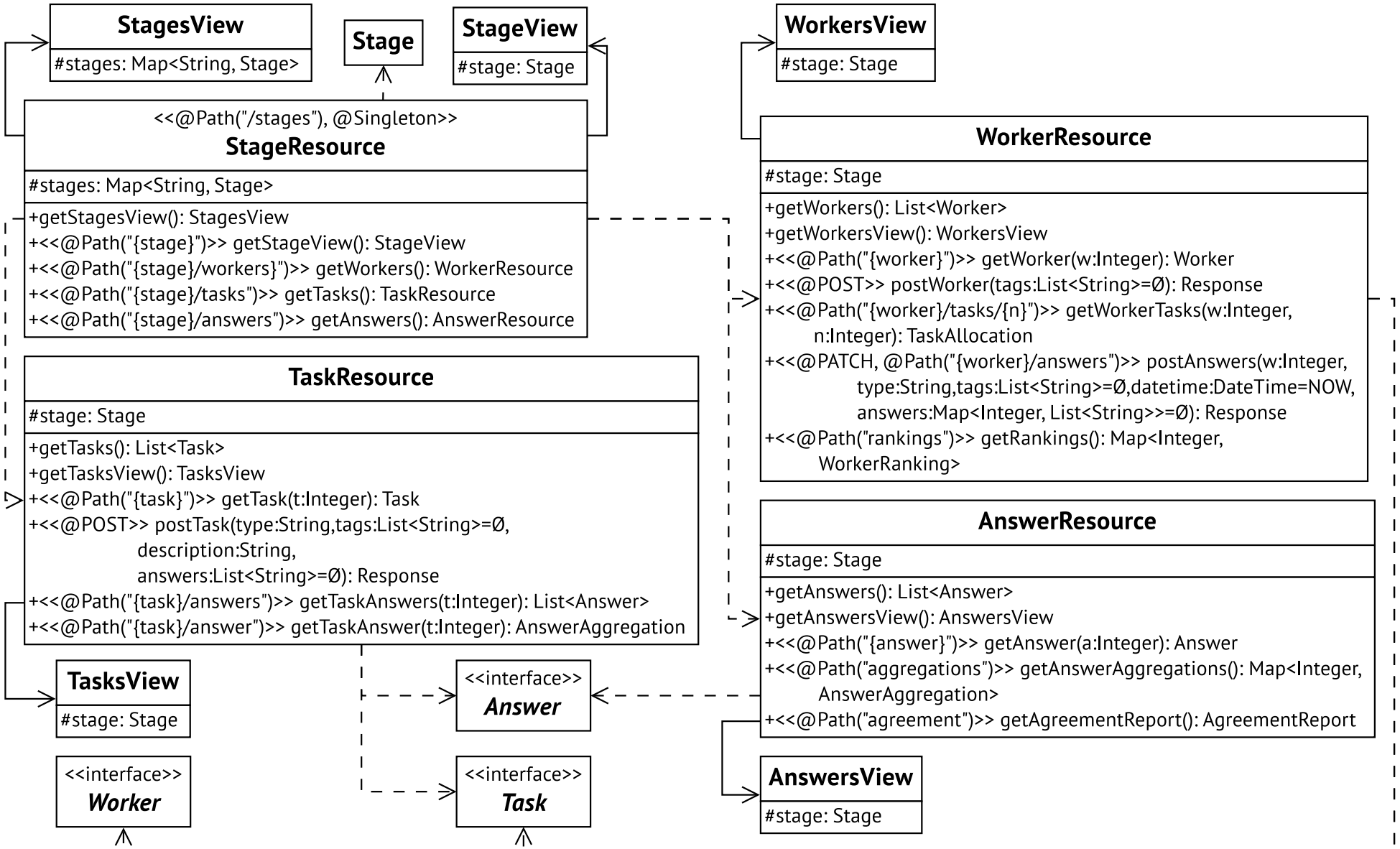
UML-диаграмма классов: процессоры



UML-диаграмма классов: сущности



UML-диаграмма классов: ресурсы



Экранные формы

Stage "gsm-genus"

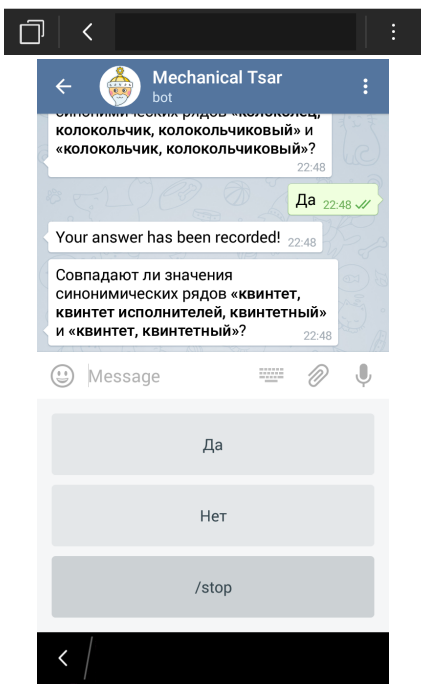
Key	Value	Action
description	Выявление вышестоящих понятий в родовидовых отношениях.	
workerCount	39	Details
workerRanker	`mtsar.processors.meta.ZenCrowd`	
taskCount	1438	Details
taskAllocator	`mtsar.processors.task.FixedNumberAllocator`	
answerCount	7357	Details
answerAggregator	`mtsar.processors.answer.MajorityVoting`	

Additional Options

Key	Value
tasksPerPage	15
answersPerTask	5

Dashboard | Stages | GitHub

Mechanical Tsar



Вычислительные и физические эксперименты

- Исследование применимости процедуры *ARC*: «добавить–удалить–подтвердить».
- Исследование применимости процедуры *GSM*: «род–вид–сопоставить».
- Построение тезауруса предметной области.

Применимость *ARC* I

Условия эксперимента

- **Данные:** 100 синсетов YARN с наибольшим количеством дубликатов по эвристике «два общих слова».

$$\exists s_1 \in S, s_2 \in S : s_1 \neq s_2 \wedge |s1 \cap s2| \geq 2 \quad (3)$$

- **Участники:** открытый вызов в VK, Facebook, Twitter.
- **Агрегация:** голос большинства.

Этап	**Участников**	**Заданий**	**Ответов**	**Длит-ть**
Добавить	24	100	501	188
Удалить	29	100	512	194
Подтвердить	4	100	300	37
Итого	36	300	1313	231

Применимость *ARC* II

- Привлечены два эксперта: ставилась оценка 1, если обработанный синсет улучшился; если нет — 0.
- Всего изменилось 84 синсета, из них улучшилось 70.

Участники		**Эксперты**	
Изменилось	84	Улучшилось	70
Не изменилось	16	Не улучшилось	14
Всего	100	**Всего**	84

Применимость *ARC* III

Оценки экспертов согласуются: различается лишь 20 оценок из 84, индекс Жаккара равен $1 - \frac{20}{84} = 74\,\%$.

Классы ошибок

- Неоднозначный синсет остался неоднозначным после обработки.
- Добавлен гипероним, гипоним или когипоним вместо синонима.
- Лишнее слово не удалено несмотря на добавление недостающих.
- Общее значение синсета изменилось.

Результаты подтверждают применимость процедуры *ARC* для уточнения лексикализации понятий.

Применимость *GSM* I

Условия эксперимента

- **Данные:** 2271 синсет YARN по тематике «безопасность жизнедеятельности», 383 кандидата-отношения.
- **Участники:** пользователи биржи TurboText.
- **Агрегация:** голос большинства, KOS, ZenCrowd.

Этап	**Участников**	**Заданий**	**Ответов**	**Длит-ть**
Род	39	1438	7357	206
Вид	18	833	4205	197
Сопоставить	18	287	1494	58
Итого	47	2558	13056	264

Применимость *GSM* II

Результаты проанализированы экспертом и сопоставлены с эвристическим методом:

$$\left|\{s' : \exists(g, s') \in \mathbb{R}\} \cap \{g' : \exists(g', s) \in \mathbb{R}\}\right| > 1. \quad (4)$$

Метод	**TP**	**TN**	**FP**	**FN**	**P**	**R**	**F_1**
Эвристика	40	102	17	128	0,70	0,24	0,36
MV	129	57	62	39	0,68	0,77	0,72
KOS	142	63	56	26	0,72	0,84	0,78
ZenCrowd	146	69	50	22	**0,74**	**0,87**	**0,80**

Применимость *GSM* III

Метод ZenCrowd продемонстрировал лучшие результаты с точки зрения точности, полноты и F_1-меры.

Классы ошибок

- Некорректное понимание лексических значений в заданиях.
- Ошибки в синсетах: чрезмерная общность или узость понятий.
- Ошибки в данных: недоверенные зашумлённые источники.

Результаты подтверждают применимость процедуры *GSM* для построения родовидовых отношений между понятиями.

Построение тезауруса предметной области

Условия эксперимента

- **Данные:** словарь предметной области, синсеты, кандидаты-отношения.
- **Участники:** пользователи биржи TurboText и CrowdFlower.
- **Агрегация:** голос большинства, KOS, ZenCrowd.

Эксперимент в процессе

Исследование выполняется в настоящее время.

Заключение

Заключение

Цели достигнуты, задачи выполнены.

Научная новизна:

- Впервые представлены краудсорсинговые процедуры утончения лексикализации понятий и построения отношений между ними.
- Впервые представлена методика коллективного построения тезауруса предметной области при помощи краудсорсинга.

Направления дальнейшей работы:

- **Интеграция с медицинскими технологиями**: использование данных ЭЭГ и других датчиков для отправки ответов.
- **Развитие модели вычислений**: асинхронное выполнение, автоматическое бюджетирование, и т. д.
- **Снижение входных барьеров**: априорная оценка сложности заданий, профилирование участников.
- Построение тезаурусов других предметных областей.

Список публикаций I

Работ в библиографической базе **Web of Science**: 2.

- *Ustalov D.* Enhancing Russian Wordnets Using the Force of the Crowd // Analysis of Images, Social Networks and Texts. — Springer International Publishing, 2014. — Vol. 436 of *Communications in Computer and Information Science*. — P. 257–264.
- *Ustalov D.* Towards Crowdsourcing and Cooperation in Linguistic Resources // Information Retrieval. — Springer International Publishing, 2015. — Vol. 505 of *Communications in Computer and Information Science*. — P. 348–358.

Список публикаций II

Работ в библиографической базе **Scopus**: 4.

- *Ustalov D., Kiselev Y.* Add-Remove-Confirm: Crowdsourcing Synset Cleansing // Application of Information and Communication Technologies (AICT), 2015 IEEE 9th International Conference on. — IEEE, 2015. — P. 143–147.
- *Ustalov D.* Crowdsourcing Synset Relations with Genus-Species-Match // Proceedings of the AINL-ISMW FRUCT. — 2015. — P. 118–124.
- *Kiselev Y., Ustalov D., Porshnev S.* Eliminating Fuzzy Duplicates in Crowdsourced Lexical Resources // Proceedings of the Eighth Global Wordnet Conference. — 2016. — P. 161–167.
- YARN: Spinning-in-Progress / P. Braslavski, D. Ustalov, M. Mukhin, Y. Kiselev // Proceedings of the Eighth Global Wordnet Conference. — 2016. — P. 58–65.

Список публикаций III

Работ в журналах перечня **ВАК**: 2 (ожидается).

- *Усталов Д.* Инструментарий краудсорсинга для механизированного труда // *Труды Института системного программирования РАН*. — 2015. — Т. 27, № 3. — С. 351–364.
- *Усталов Д.* Коллективные потоковые вычисления: реляционные модели и алгоритмы // На рецензировании.

Список публикаций IV

Прочие работы и публикации: 3.

- *Ustalov D.* Teleboyarin—Mechanized Labor for Telegram // Proceedings of the AINL-ISMW FRUCT. — 2015. — P. 195–197.
- *Усталов Д.* Свидетельство Роспатента о государственной регистрации программы для ЭВМ «Адаптивная система управления процессом краудсорсинга» № 2015662640 от 30.11.2015.
- *Усталов Д.* Свидетельство Роспатента о государственной регистрации программы для ЭВМ «Система автоматизации процесса коллективного построения баз данных» № 2015662780 от 01.12.2015.

Апробация работы I

- 14-я конференция европейского отделения Ассоциации по компьютерной лингвистике **EACL 2014** (г. Гётеборг, Швеция).
- Международная суперкомпьютерная конференция «**Научный сервис в сети Интернет**: многообразие суперкомпьютерных миров» (г. Новороссийск, 2014 г.)
- 14-я национальная конференция по искусственному интеллекту **КИИ-2014** (г. Казань).
- 5-я международная конференция по инженерии знаний и Семантической паутине **KESW 2014** (г. Казань).
- 16-я всероссийская научная конференция **RCDL 2014** (г. Дубна).
- 3-я и 4-я международная конференция по анализу изображений, социальных сетей и текстов **АИСТ'2014** и **АИСТ'2015** (г. Екатеринбург).

Апробация работы II

- 9-й весенне-летний коллоквиум молодых учёных по программной инженерии **SYRCoSE 2015** (г. Самара).
- 21-я международная конференция по компьютерной лингвистике «**Диалог 2015**» (г. Москва).
- 9-я международная конференция по использованию информационно-коммуникационных технологий **AICT2015** (г. Ростов-на-Дону).
- Международная конференция **AINL-ISMW FRUCT** (г. Санкт-Петербург, 2015 г.)
- 8-я глобальная конференция по ворднетам **GWC 2016** (г. Бухарест, Румыния).
- Международная конференция «**Современные проблемы математики и её приложений**» (г. Екатеринбург, 2016 г.)
- 186-е заседание семинара **московской секции ACM SIGMOD** (г. Москва, 2016 г.)

Поддержка научными фондами

- Исследование выполнено при финансовой поддержке РГНФ: проект «Новый открытый электронный тезаурус русского языка» № 13-04-12020 и проект «Интеграция тезаурусов RussNet и YARN» № 16-04-12019.
- Поддержка данного проекта осуществлена в рамках благотворительной деятельности, на средства, предоставленные Фондом Михаила Прохорова.
- Работа выполнена при финансовой поддержке стипендии Президента Российской Федерации молодым учёным и аспирантам № СП-773.2015.5.
- Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 16-37-00354 мол_а «Методы автоматизации процесса коллективного построения лингвистических ресурсов».

Спасибо за внимание!

Дмитрий Усталов

https://linkedin.com/in/ustalov

http://ustalov.imm.uran.ru/

dau@imm.uran.ru