Методы повышения эффективности процесса коллективного построения лексических ресурсов

Д. А. Усталов

ИММ УрО РАН

31 марта 2016 г.

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.

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.

Построение языковых ресурсов

- **Традиционный** подход — построение ресурсов узким коллективом экспертов-лексикографов. Процесс оказывается длительным и дорогостоящим.
  Пример: WordNet.
- **Автоматический** подход — построение производного ресурса на основе существующих путём автоматического сопоставления. Необходимы качественные данные.
  Пример: BabelNet.
- **Автоматизированный** подход — построение языковых ресурсов при помощи краудсорсинга с применением методов обеспечения качества.
  Пример: OpenCorpora.

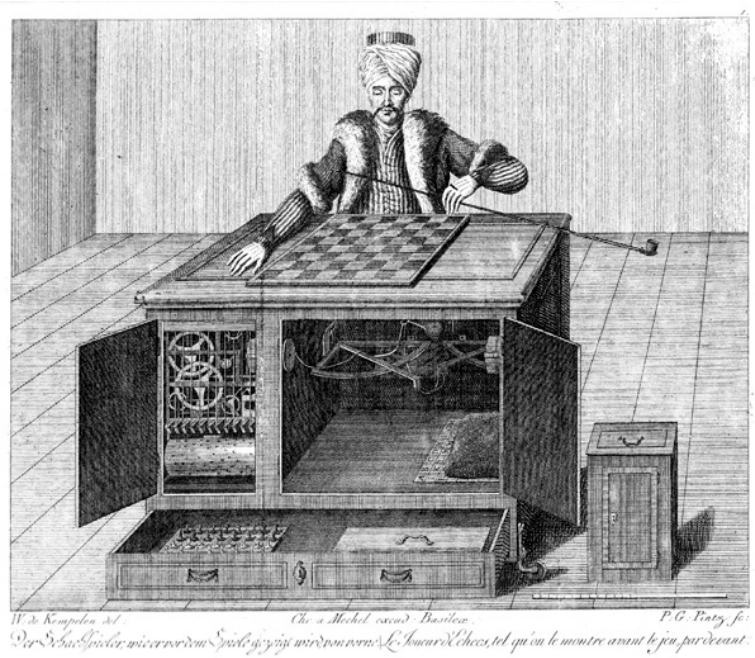
Тезаурусы русского языка

- **RussNet**: 30 тыс. синсетов, 45 тыс. отношений. Данные не представлены в едином формате и не опубликованы.
- **РуТез**: 55 тыс. понятий, 210 тыс. отношений. Успешный ресурс с ограничительной лицензией.
- Автоматический перевод WordNet: **WordNet.ru** (31 тыс. синсетов) и **Russian Wordnet** (недоступен).
- **Викисловарь** (русский): 190 тыс. слов, 70 тыс. отношений. Отношения между словами, оценка качества не проводилась.
- **UNL**: 62 тыс. универсальных слов, 90 тыс. связей между ними. Оценка качества не проводилась.
- **BabelNet** (русский): 2,5 млн. понятий, 380 млн. отношений (всего). Ограничительная лицензия, оценка качества не проводилась.

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*.

Биржи краудсорсинга

Определение

Биржа краудсорсинга *— человеко-машинная платформа для размещения и выполнения микрозадач.*

Биржа	**Цена**	**Кол-во задач**	**Заказчик?**	**Участник?**
MTurk	$0.60	3050	Нет*	Да
CrowdFlower	$0.05	47	Да	Да
microWorkers	$0.65	93	Да	Да
Яндекс.Толока	$0.01	3	Да*	Да
TurboText	$0.06	289	Да	Да

Данные от 1 октября 2015 г.

Обзор литературы

- Методики решения задач.
- Методы обеспечения качества.
- Программное обеспечение.
- Подходы к оценке качества языковых ресурсов.

Методики решения задач I

Исследователи разрабатывают специализированные методики для решения задач при помощи краудсорсинга.

- **ESP Game** (2006) — игрофицированная разметка изображений.
- **Find-Fix-Verify** (2010) — вычитка, перефразирование и улучшение документов Microsoft Word.
- **PlateMate** (2011) — оценка калорийности блюд по фотографиям.
- **CrowdER** (2012) — обнаружение дубликатов и связывание записей о товарах.
- **TWSI** (2013) — построение языковых ресурсов на основе явления лексической замещаемости.

Методики решения задач II

Предпринимаются попытки адаптации популярных моделей программирования к области краудсорсинга микрозадачами.

- **CrowdForge** (2011) — адаптация модели вычислений MapReduce.
- **CrowdDB** (2011) — оператор `CROWD` для размещения задач при выполнении SQL-запроса.
- **Qurk** (2011) — SQL-подобный язык для формулирования и выполнения задач.
- **Turkomatic** (2011) — формулирование цепочек заданий на естественном языке.
- **CrowdWeaver** (2012) — потоковая модель вычислений для краудсорсинга.

Методы обеспечения качества

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

- Метод **Давина-Скина** (1979) — построение матриц ошибок при помощи EM-алгоритма.
- **GLAD** (2009) — вывод сложности заданий, квалификации участников и правильных ответов на основе EM-алгоритма.
- Алгоритм **Каргера-Оха-Шаха** (2011) — оптимальное по порядку назначение заданий и вывод правильных ответов.
- **ZenCrowd** (2012) — вывод квалификации участников и правильных ответов 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 языковых ресурсов I

- **Золотой стандарт** — автоматическое сравнение ресурса с аналогичным ресурсом известного качества.
  Пример: WordNet.
- **Экспертная оценка** — ручной анализ ресурса с привлечением эксперта.
  Пример: OpenCorpora.
- **Разделённая задача** — соревнование («дорожка») по применению ресурса для решения поставленной прикладной задачи.
  Пример: RUSSE.

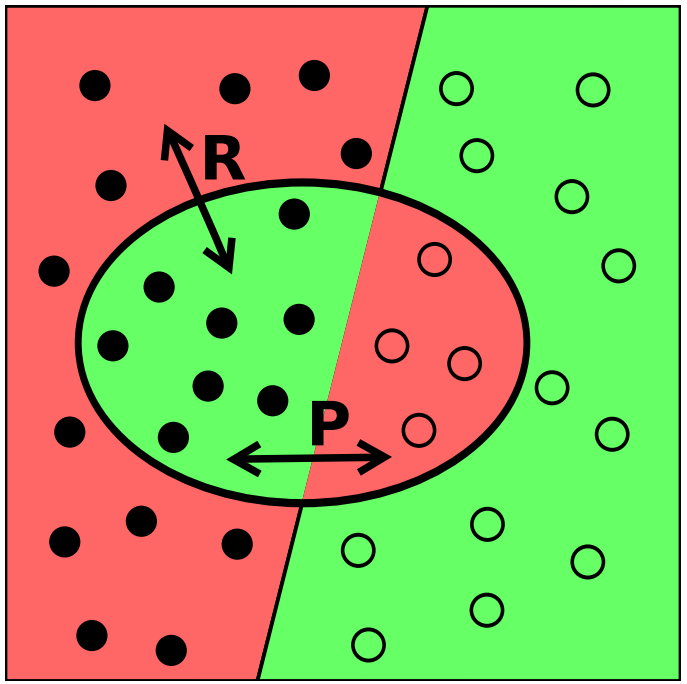
Подходы к оценке качества языковых ресурсов II

Ответы:

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

Метрики:

- 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.

Формулировка исследуемой задачи

Разработать методы повышения эффективности процесса построения лексических ресурсов путём обработки существующих словарей и тезаурусов неизвестного качества при помощи краудсорсинга.

Выводы по первой главе

- Определены понятия языкового ресурса и краудсорсинга.
- Рассмотрено их практическое применение.
- Сформулирована исследуемая задача.

Методика построения тезауруса предметной области I

Постановка задачи

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

Методика построения тезауруса предметной области II

1. Извлечь из $\mathbb{S}$ все понятия, содержащие слова из $\mathbb{D}$.
2. Извлечь из $\mathbb{R}$ все пары, слова которых представлены в синсетах.
3. Построить неоднозначные родовидовые отношения при помощи краудсорсинга.
4. Выполнить уточнение лексикализации понятий при помощи краудсорсинга.
5. Объединить дубликаты-когипонимы автоматически.

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 упаковывания и распаковывания.
- 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.*

$$W = (S \cup Y \cup D, E) \tag{1}$$

Схема коллективных вычислений II

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: $\mathrm{PK}(v) \subseteq H(v)$;
- множество входящих вершин в вершину v: $In(v) \subset V$.

Схема коллективных вычислений III

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:* $In(d) = \varnothing$.

Согласованные схемы коллективных вычислений

Этапы синхронизации и средства предварительной обработки данных реализуются в виде программного обеспечения различной сложности, что вносит дополнительные требования к детерминированности описания методики краудсорсинга.

Определение

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

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

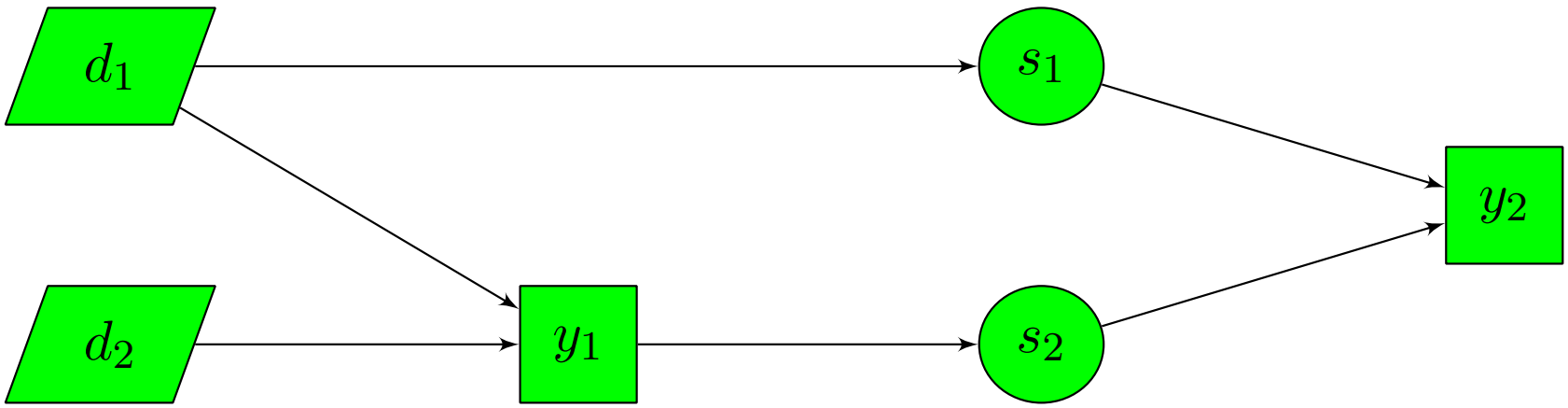
Выполнение схем коллективных вычислений

Алгоритм 1 Синхронный алгоритм выполнения

Require: $V = S \cup Y \cup D, |D| > 0$

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

Ensure: $(\forall v \in V)(M_v = \textbf{true})$


d_1
d_2
y_1
s_1
s_2
y_2

Уточнение лексикализации понятий I

Определение

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

Пример

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

Синсеты могут быть не идеальны.

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

Уточнение лексикализации понятий II

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.

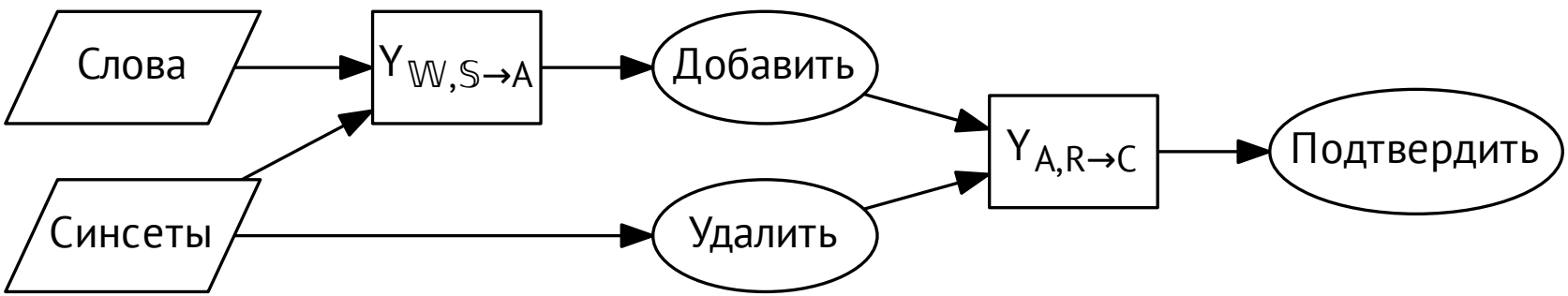
Процедура «добавить–удалить–подтвердить» I

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

Процедура «добавить–удалить–подтвердить» III

Определение

$$S_{ARC} = \{A, R, C\} \tag{3.1}$$

$$Y_{ARC} = \{Y_{\mathbb{W},\mathbb{S}\to A}, Y_{A,R\to C}\} \tag{3.2}$$

$$D_{ARC} = \{\mathbb{W}, \mathbb{S}\} \tag{3.3}$$

$$\begin{aligned} E_{ARC} = \ & \big\{(\mathbb{S}, Y_{\mathbb{W},\mathbb{S}\to A}), (\mathbb{W}, Y_{\mathbb{W},\mathbb{S}\to A}), (\mathbb{S}, R), \\ & (Y_{\mathbb{W},\mathbb{S}\to A}, A), (A, Y_{A,R\to C}), \\ & (R, Y_{A,R\to C}), (Y_{A,R\to C}, C)\big\} \end{aligned} \tag{3.4}$$

$$ARC = \left(S_{ARC} \cup Y_{ARC} \cup D_{ARC}, E_{ARC}\right) \tag{3.5}$$

Построение родовидовых отношений I

Определение

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

Пример

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

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

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

Построение родовидовых отношений II

Постановка задачи

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

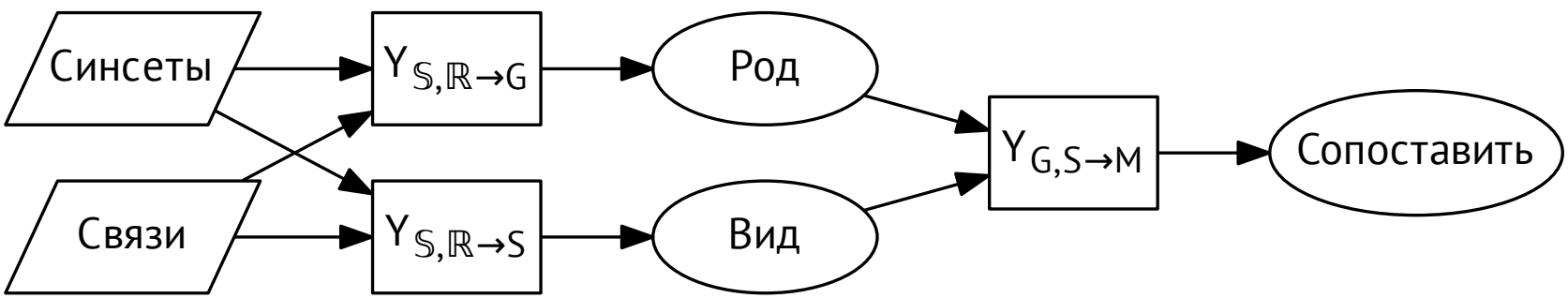
Процедура «род–вид–сопоставить» I

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	$H(M) = \{G.id, S.id, Y_{G,S\to M}.s_1, Y_{G,S\to M}.s_2, (b, \texttt{BOOL})\}$

Процедура «род–вид–сопоставить» III

Определение

$$S_{GSM} = \{G, S, M\} \tag{4.1}$$

$$Y_{GSM} = \{Y_{\mathbb{S},\mathbb{R}\to G}, Y_{\mathbb{S},\mathbb{R}\to S}, Y_{G,S\to M}\} \tag{4.2}$$

$$D_{GSM} = \{\mathbb{S}, \mathbb{R}\} \tag{4.3}$$

$$\begin{aligned} E_{GSM} = \ & \big\{(\mathbb{S}, Y_{\mathbb{S},\mathbb{R}\to G}), (\mathbb{S}, Y_{\mathbb{S},\mathbb{R}\to S}), \\ & (\mathbb{R}, Y_{\mathbb{S},\mathbb{R}\to G}), (\mathbb{R}, Y_{\mathbb{S},\mathbb{R}\to S}), \\ & (Y_{\mathbb{S},\mathbb{R}\to G}, G), (Y_{\mathbb{S},\mathbb{R}\to S}, S), \\ & (S, Y_{G,S\to M}), (G, Y_{G,S\to M}), \\ & (Y_{G,S\to M}, M)\big\} \end{aligned} \tag{4.4}$$

$$GSM = \big(S_{GSM} \cup Y_{GSM} \cup D_{GSM}, E_{GSM}\big) \tag{4.5}$$

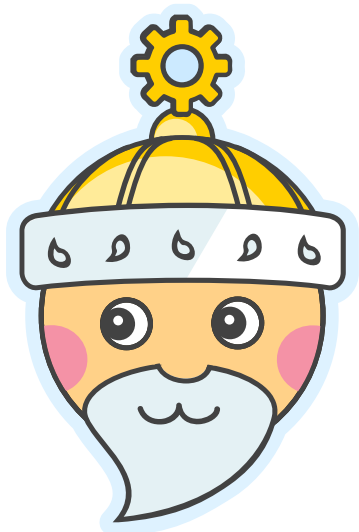
Выводы по второй главе

- Предложена методика коллективного построения тезауруса предметной области.
- Предложена модель коллективных потоковых вычислений.
- Предложена процедура «добавить–удалить–подтвердить» для уточнения лексикализации понятий.
- Предложена процедура «род–вид–сопоставить» для построения родовидовых отношений между понятиями.

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/

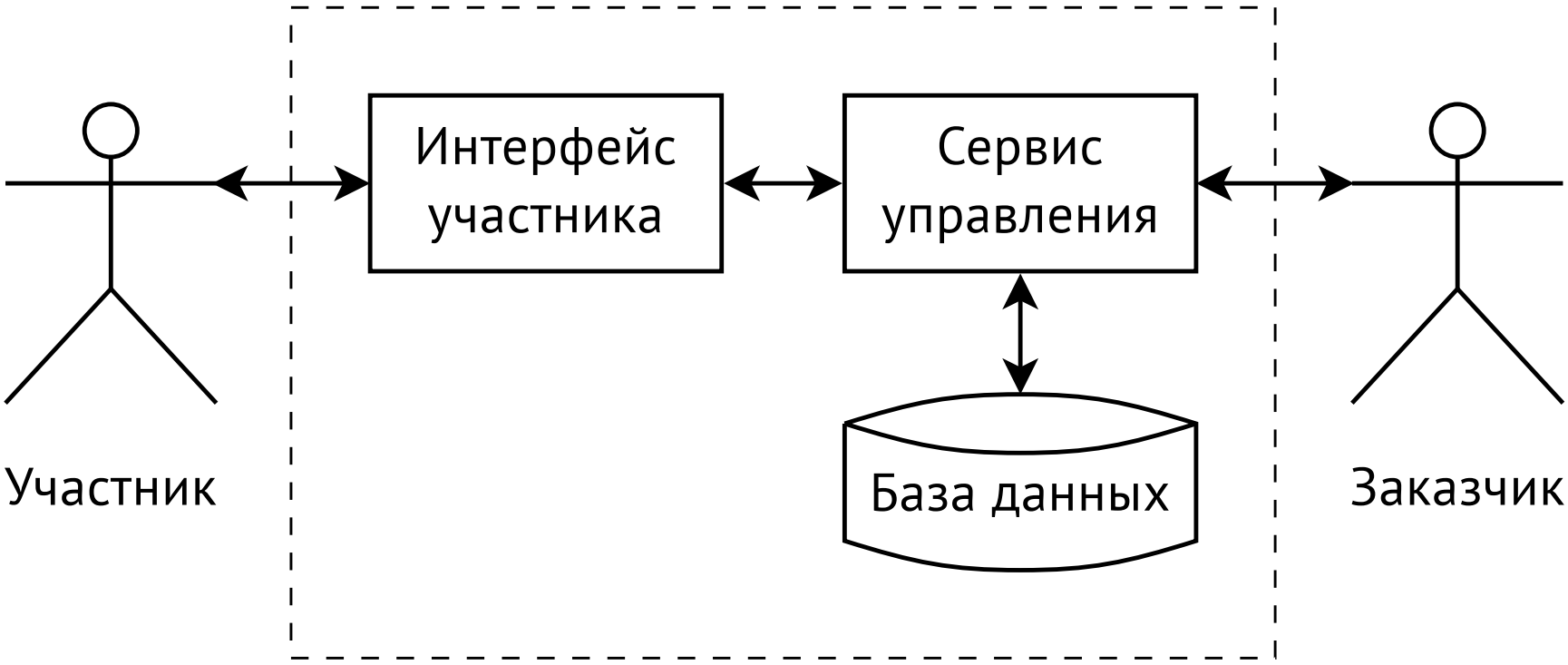
Функциональность сервиса управления

- Назначение заданий участникам.
- Приём ответов от участников.
- Агрегация ответов на задания.
- Оценка согласованности ответов.
- Ввод и вывод данных заказчика.

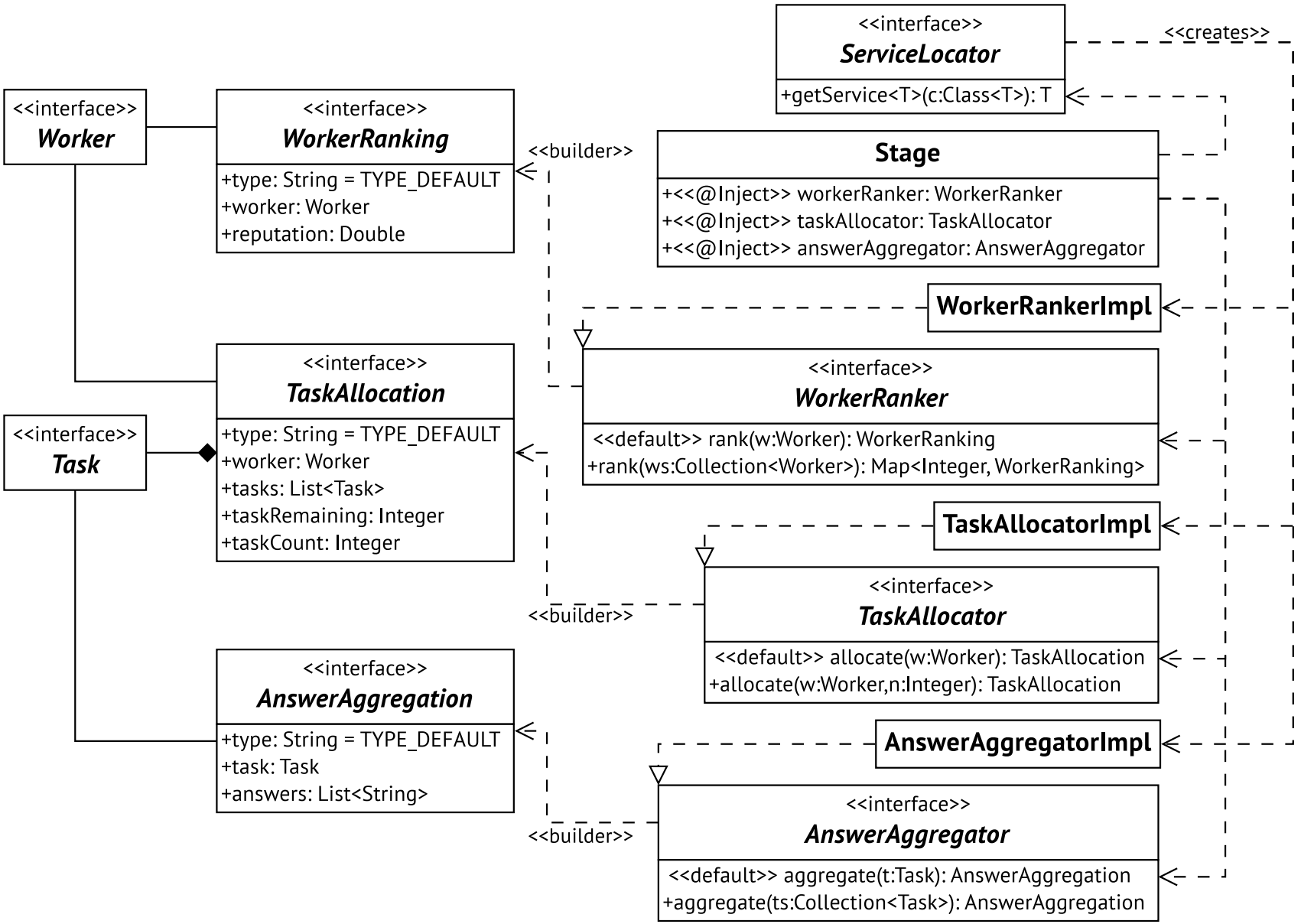
Функциональность интерфейса участника

- Запрос регистрации участника.
- Запрос назначения задания.
- Запрос приёма ответа.

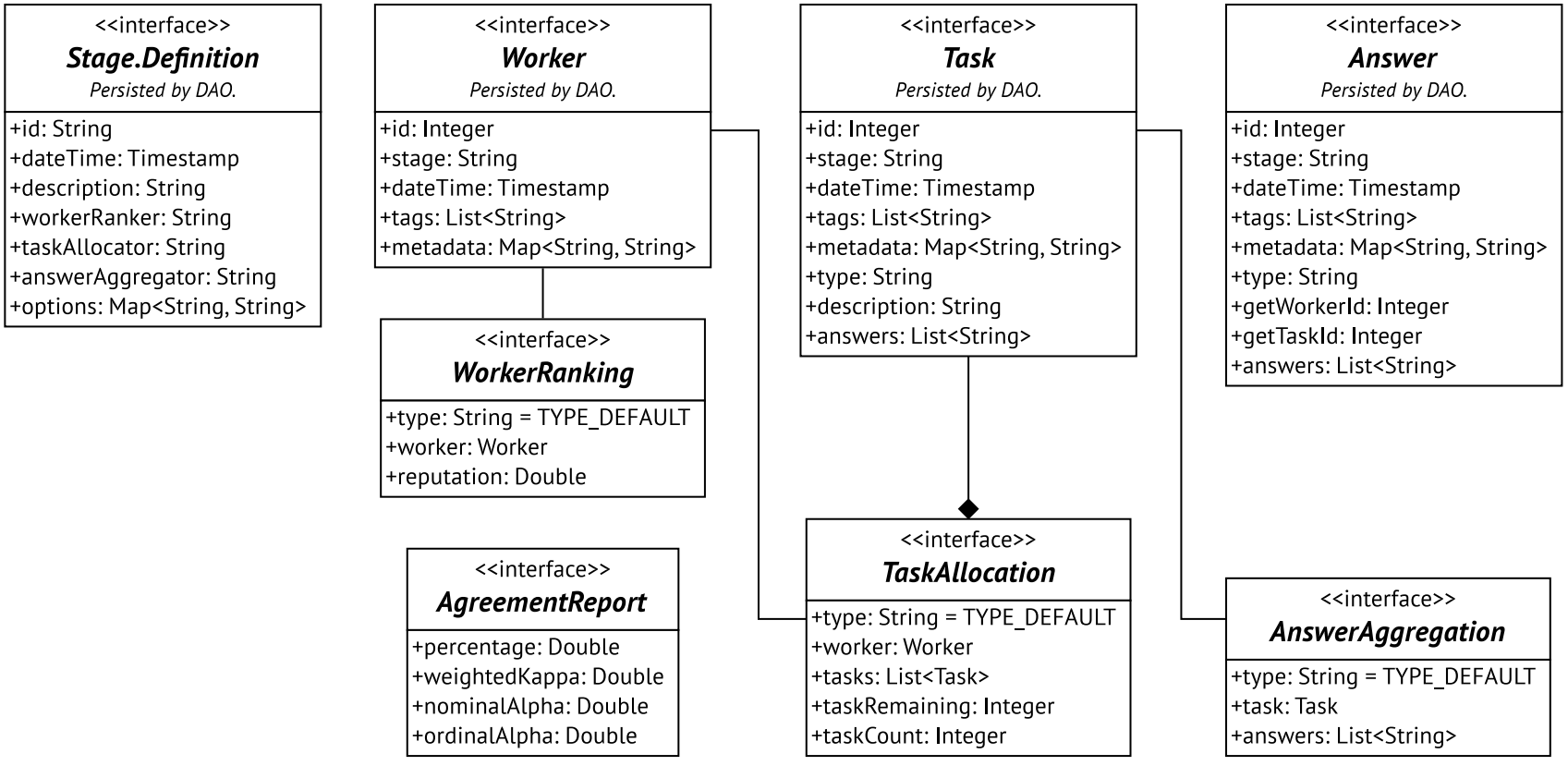
Архитектура системы



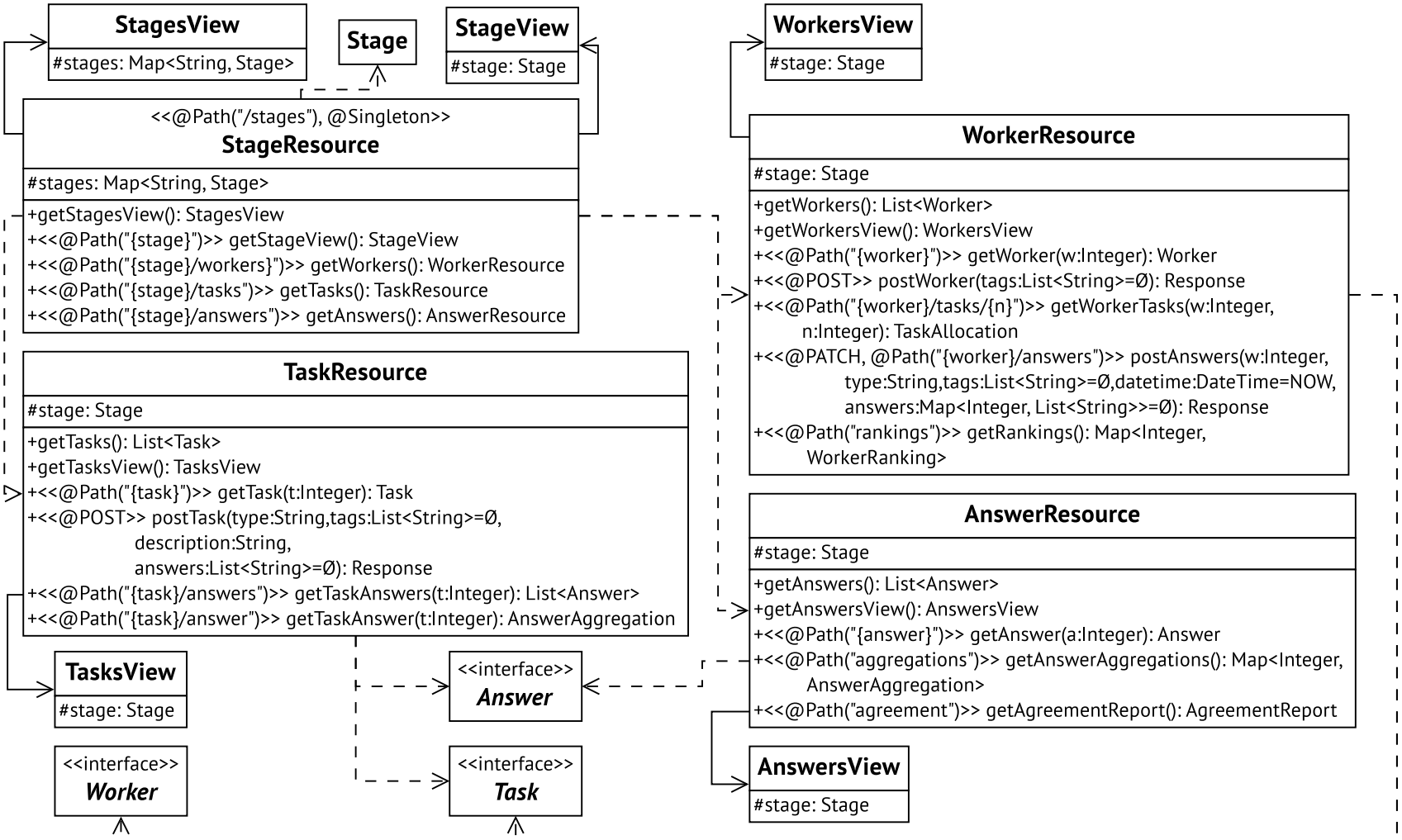
UML-диаграмма классов: процессоры



UML-диаграмма классов: сущности



UML-диаграмма классов: ресурсы



Сервис управления: сводка

Dashboard

Key	Value
Version	0.0.1-SNAPSHOT
Java	1.8.0_45-internal-b14
Processes	3
Workers	73
Tasks	300
Answers	1313

Dashboard | Processes | GitHub

Mechanical Tsar

Сервис управления: этапы разметки

Processes

ID	Description	Workers	Tasks	Answers
arc-add	Add-Remove-Confirm: Add.	24	100	501
arc-rm	Add-Remove-Confirm: Remove.	29	100	512
arc-confirm	Add-Remove-Confirm: Confirm.	20	100	300

Dashboard | Processes | GitHub

Mechanical Tsar

Сервис управления: состояние этапа разметки

Process "arc-add"

Key	Value	Action
description	Add-Remove-Confirm: Add.	
workerCount	24	Details
workerRanker	`mtsar.processors.worker.ZeroRanker`	
taskCount	100	Details
taskAllocator	`mtsar.processors.task.FixedNumberAllocator`	
answerCount	501	Details
answerAggregator	`mtsar.processors.answer.EmptyAggregator`	

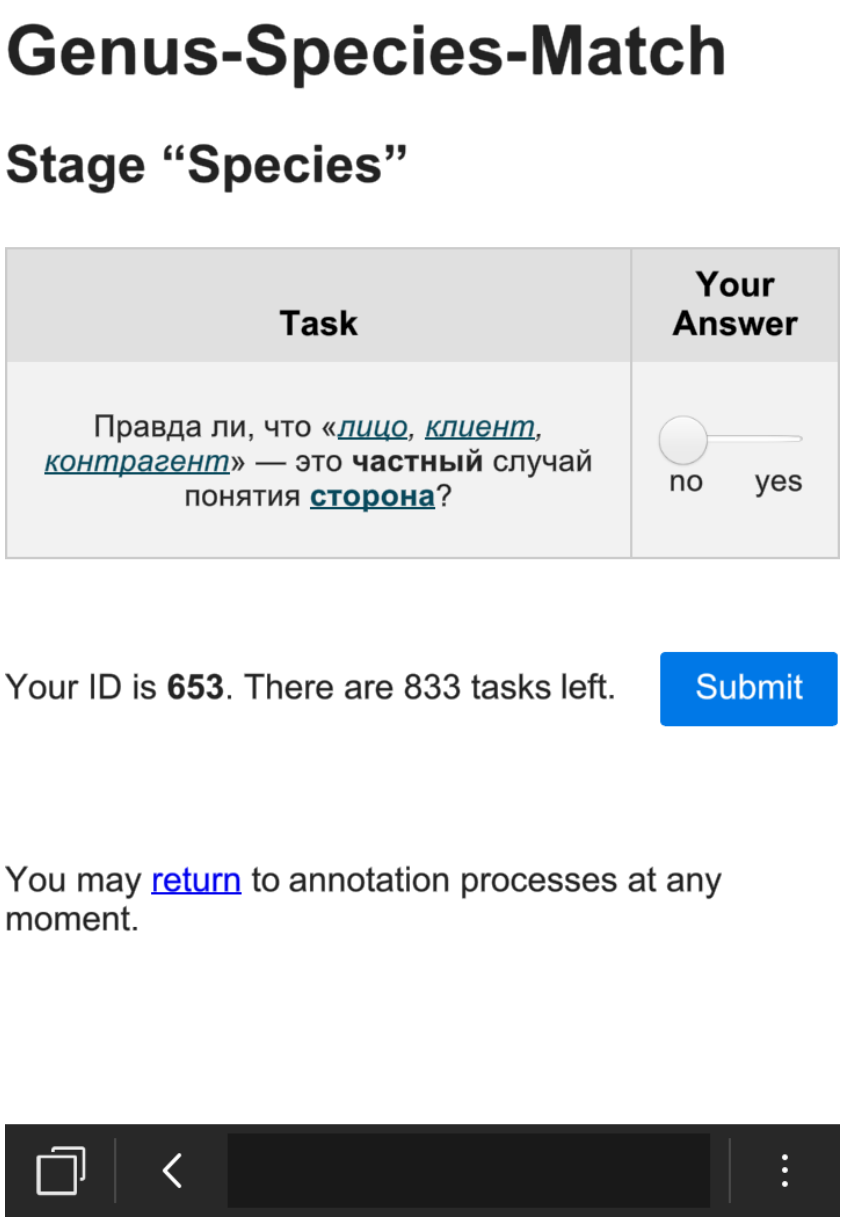
Additional Options

Key	Value
answersPerTask	5

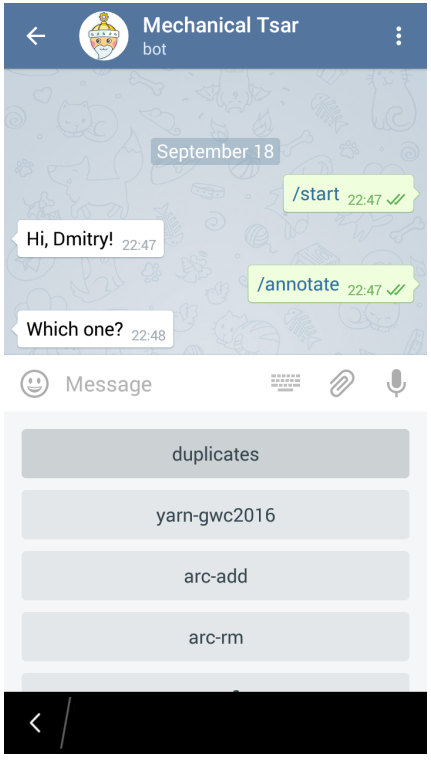
Dashboard | Processes | GitHub

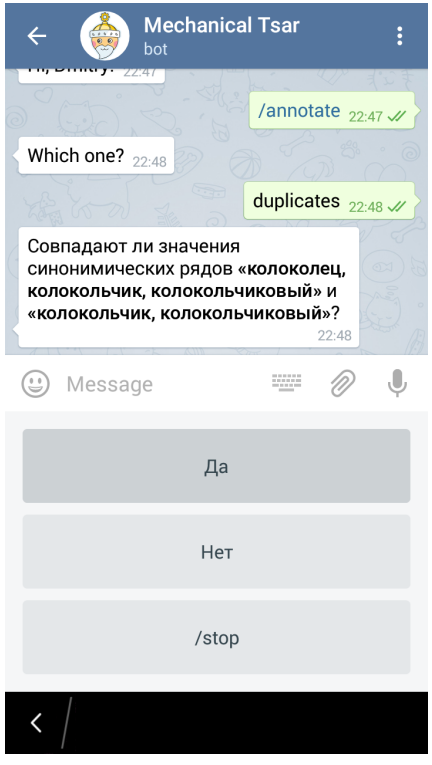
Mechanical Tsar

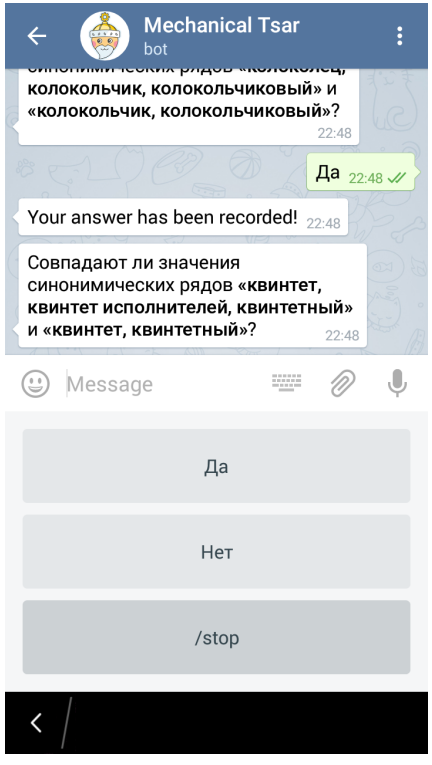
Веб-интерфейс участника

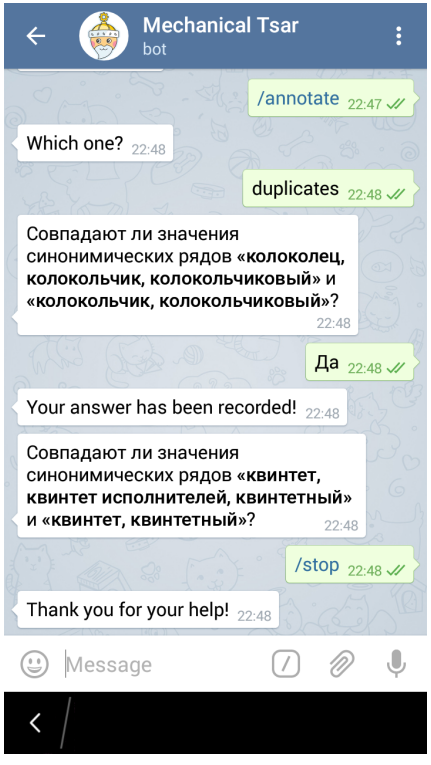


Telegram-интерфейс участника









Выводы по третьей главе

- Разработан комплекс программ для управления процессом краудсорсинга.
- Комплекс программ поддерживает разметку с разных устройств и платформ.

Вычислительные и физические эксперименты

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

Применимость ARC I

Условия эксперимента

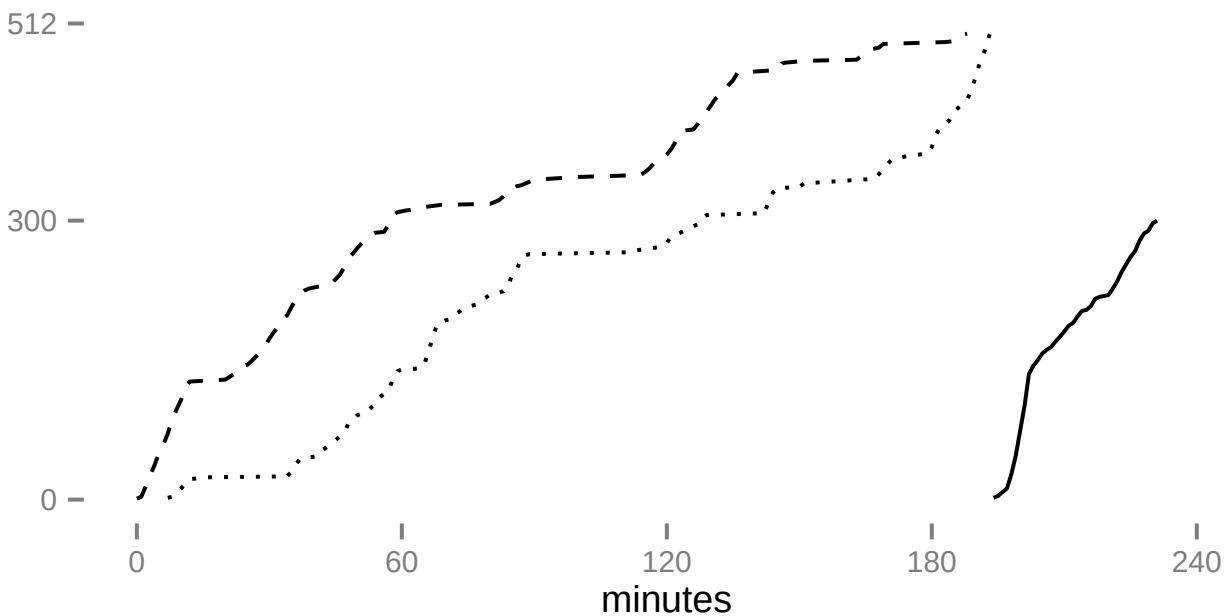
- **Данные:** 100 синсетов YARN с наибольшим количеством дубликатов по эвристике «два общих слова».

$$\exists s_1 \in S, s_2 \in S : s_1 \neq s_2 \wedge |s1 \cap s2| \geq 2 \tag{5}$$

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

Применимость ARC II

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 III

- 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 IV

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.

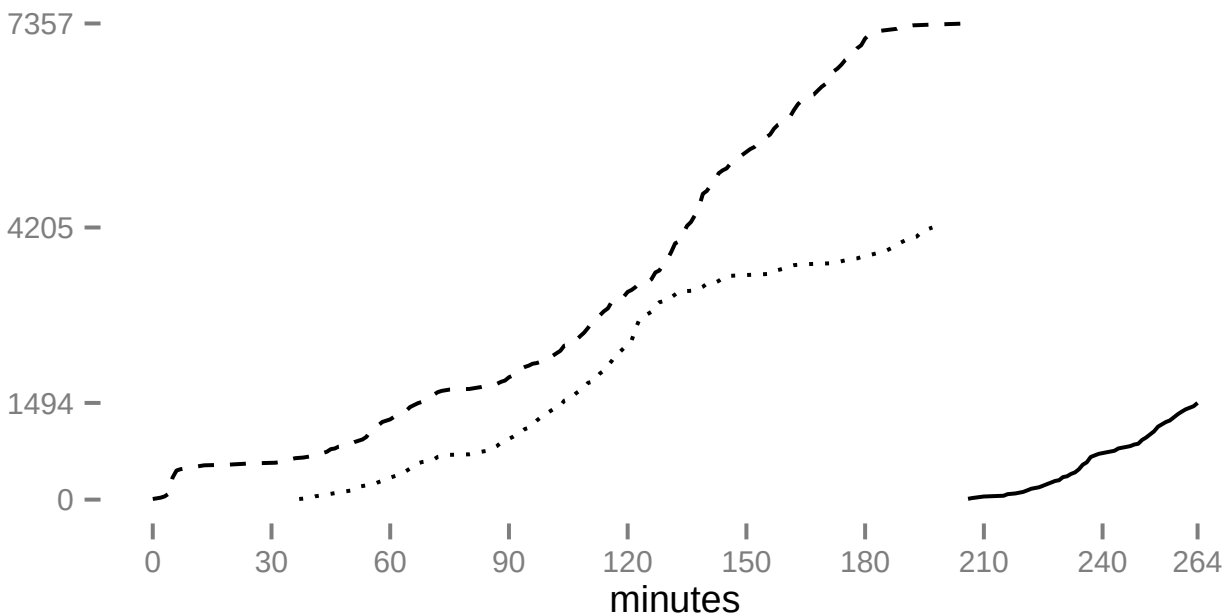
Применимость GSM I

Условия эксперимента

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

Применимость GSM II

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 III

Результаты проанализированы экспертом и сопоставлены с эвристическим методом:

$$\left|\{s' : \exists(g, s') \in \mathbb{R}\} \cap \{g' : \exists(g', s) \in \mathbb{R}\}\right| > 1. \quad (6)$$

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 IV

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 I

Условия эксперимента

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

Построение тезауруса предметной области II

Эксперимент в процессе

Исследование выполняется в настоящее время.

Выводы по четвёртой главе

- Подтверждена применимость процедуры «добавить–удалить–подтвердить».
- Подтверждена применимость процедуры «род–вид–сопоставить».
- Построен русскоязычный тезаурус предметной области.

Заключение

Заключение

Цели достигнуты, задачи выполнены.

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.

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).

Апробация работы II

- 3-я и 4-я международная конференция по анализу изображений, социальных сетей и текстов **АИСТ'2014** и **АИСТ'2015** (г. Екатеринбург).
- 9-й весенне-летний коллоквиум молодых ученых 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.)

Апробация работы III

- 8-я глобальная конференция по ворднетам **GWC 2016** (г. Бухарест, Румыния).
- Международная молодёжная школа-конференция «**Современные проблемы математики и её приложений**» (г. Екатеринбург, 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!

Вопросы?

Дмитрий Усталов

https://linkedin.com/in/ustalov

https://ustalov.name/

dau@imm.uran.ru