«Московский авиационный институт»
(Национальный исследовательский университет)

«Информатика: проблемы, методы, технологии» (IPMT-2022)

Исследование инструментов морфологического анализа текстов на русском языке для повышения точности алгоритмов обработки в библиотеке JMorfSdk

к.т.н., доцент каф. 319 **Полицына Е.В.**,
к.т.н., доцент каф. 319 **Полицын С.А.**,
аспирант, каф. 319 **Поречный А.С.**,
студент, каф. 319 **Рыкунов А.Н.**

Применение компьютерной лингвистики

- Распознавание речи
- Электронная почта
- Информационный поиск
- Определение тональности текста
- Перевод текста
- Грамматика и проверка орфографии
- Реферирование текста
- Обслуживание клиентов (чат-боты)

Google

Морфологический анализ

Пример получения характеристик для слов предложения:

У ног толкуют волны. (А. Новиков-Прибой)

у : (предлог) у;

нога : (имя существительное) ног : морф. характеристики – род. падеж, мн. ч., жен. р.;

толковать : (глагол) толкуют : морф. характеристики – наст. вр., мн. ч., 3-е лицо;

волна : (имя существительное) волны : морф. характеристики – им. падеж, ед. ч, жен. р.

JMorfSdk

- ✔ Модуль **морфологического анализа** в составе **фреймворка TAWT**
- ✔ Получение **морфологических характеристик** слова
- ✔ Получение **начальной формы** слова
- ✔ **Генерация слов** по заданным характеристикам

1. Использует **OpenCorpora** (360 тысяч уникальных слов, более 5 млн. словоформ).
2. Представление **морфологических характеристик** слов в виде **бинарной шкалы.**
3. **Целочисленное** внутреннее **представление слов**. Для исключения коллизий используется два алгоритма хэширования.
4. **Строковое представление слов** хранится в **базе данных.**

Инструменты морфологического анализа

TreeTagger
Python, Java, Ruby

pymorphy2
Python, Java

Natasha
Python

TAWT
Java

AoT
Java

Pullenti
C#, Java, Python

Russian Morphology for Lucene
Java

Параметры сравнения инструментов

- Количество найденных слов – процент найденных слов от общего количества слов в корпус
- Количество верно полученных начальной формы – процент верно определенных начальных форм слов от общего количества слов в корпусе
- Время получения начальной формы – 95-й процентиль
- Количество верно определенных морфологических характеристик:
  - Без учета снятия омонимии – процент верно определенных морфологических характеристик без учета снятия омонимии от количества слов, найденных в словаре анализатора;
  - С учетом снятия омонимии – процент верно определенных морфологических характеристик с учетом снятия омонимии от количества слов, найденных в словаре анализатора.

Национальный корпус русского языка

- Корпус со снятой омонимией на русском языке:
  - 95 056 предложений;
  - 1 023 297 слов.
- Тексты из различных областей и написанных разным стилем:
  - стили: художественный, публицистический, научный и др.;
  - области: наука, искусство, политика, спорт, медицина и др.

Результаты сравнения работы морфологических анализаторов

Название инструмента	Количество найденных слов, %	Количество верно полученных начальных форм, %	95-й процентиль времени получения начальной формы, мс
JMorfSdk	95,6	82,3	0,001488
TreeTagger	99,0	92,3	0,269281
PullEnti	95,7	82,5	0,022700
pymorphy2	98,4	95,0	0,025552
RussianMorphology	94,8	90,4	0,001932
AoT	95,5	90,5	0,001644
Natasha	95,6	94,6	0,275177

Результаты сравнения работы морфологических анализаторов

Название инструмента	Количество верно определенных морфологических характеристик, %	
	Без учета снятия омонимии	С учетом снятия омонимии
JMorfSdk	90,8	70,2
TreeTagger	76,2	76,2
PullEnti	63,3	48,4
pymorphy2	91,7	76,6
RussianMorphology	67,6	63,0
AoT	96,8	72,8
Natasha	82,8	82,8

Преимущества библиотеки JMorfSdk

- Скорость работы
- Возможность генерации слов по заданным морфологическим характеристикам
- Реализация на распространенном языке программирования для промышленной разработки ПО
- Соответствие современным требованиям к Java-библиотекам

Недостатки библиотеки JMorfSdk

- Размер словарь (95,6% найденных слов)
- Особенности начальных форм глаголов (начальной формой является форма 1-го лица, единственного числа “сказал” – “скажу”)
- Невозможность получения словоформ с “ё”, при написании слов через “е” (“ещё” – “еще”)
- Распознавание чисел не во всех формах написания: на данный момент распознаются только целочисленные значения (55, 10000)
- Нет разрешения омонимии

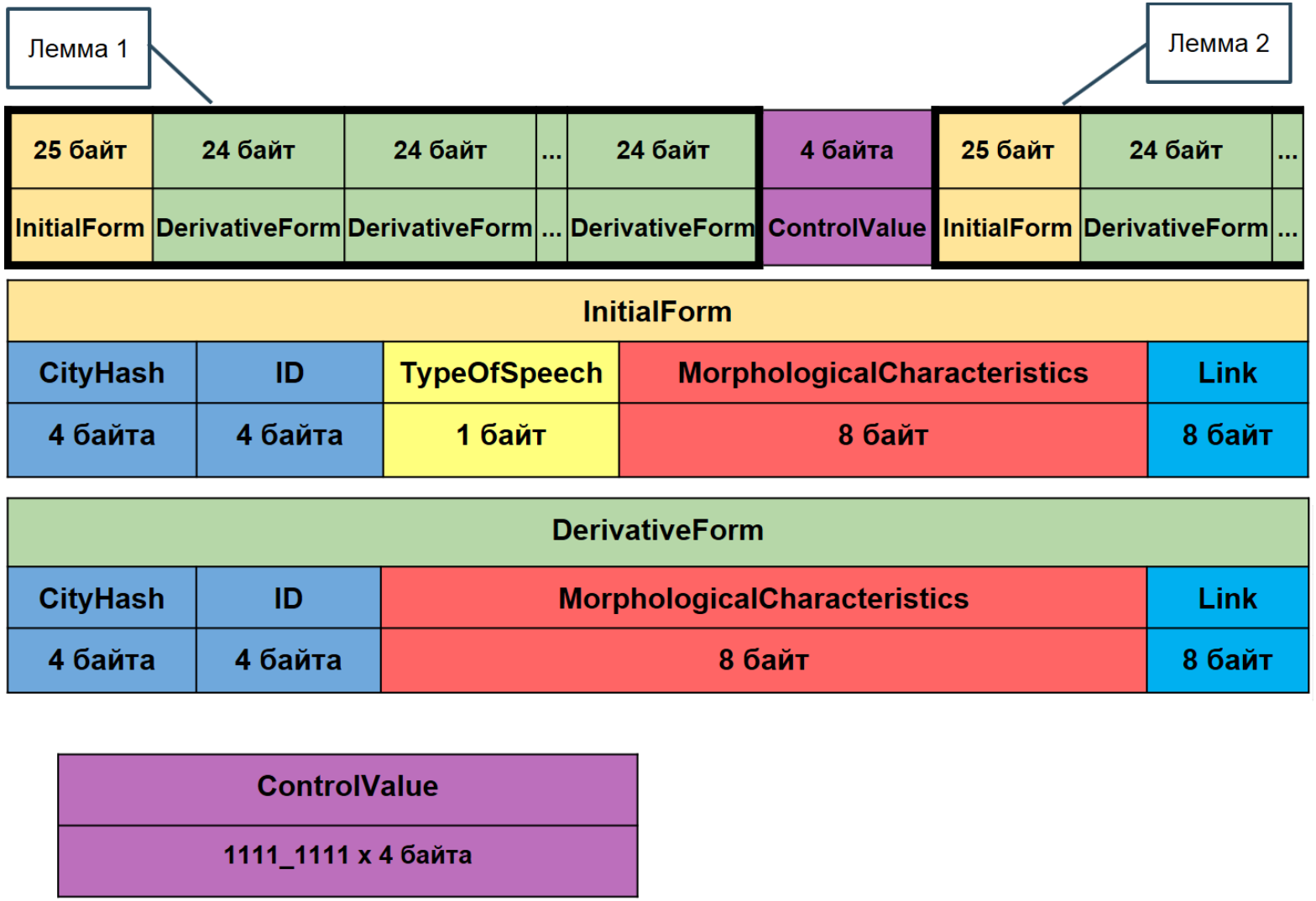
Направления развития JMorfSdk

- Обновление текущей версии словаря (391 778 лемм, 5 140 595 словоформ)
- Добавление отдельного словаря отсутствующих слов
- Добавление инфинитива в качестве начальной формы глаголов
- Добавление в словарь словоформ с “е” с сохранением связей между ними и словоформами с “ё”
- Расширение набора распознаваемых форматов данных: дробные числа (5,3; 5.3), даты (YYYY-MM-DD, YYYY-M-DD, …), номера телефонов и др.
- Добавление частичного разрешения омонимии

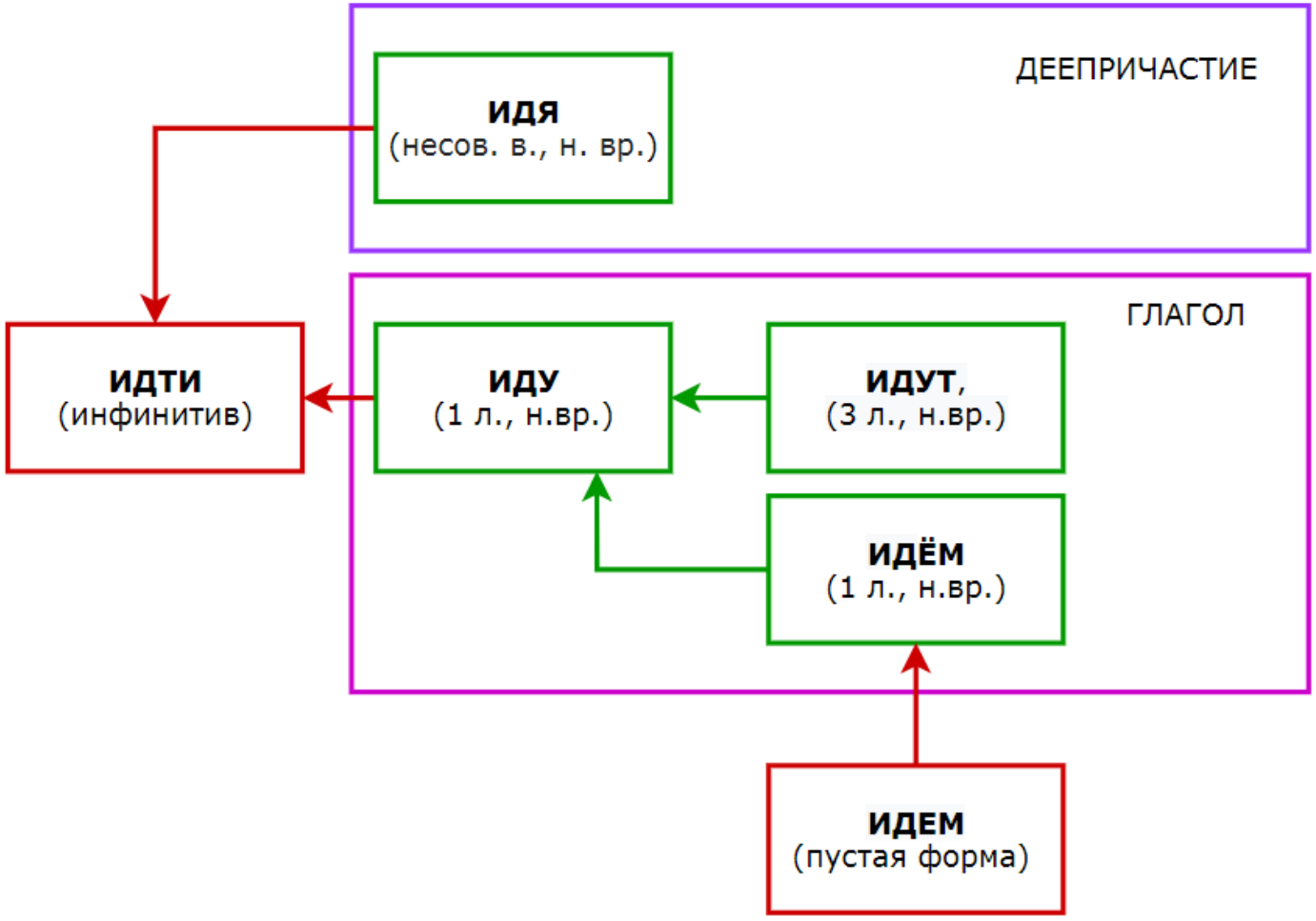
Добавление в словарь отсутствующих слов

- Анализ большого количества текстов:
  - **Твиттер:** 11 990 077 сообщений – **103 478 702** слов, **67 720** неизвестных слов, употреблённых более 10 раз
  - **lenta.ru:** 800 975 новостей – **138 946 407** слов, **442 407** неизвестных слов
  - **Корпус текстов**: 92 731 текст – **424 933 881** слов, **2 034 257** неизвестных слов
- Результаты предварительного анализа:
  - Опечатки (пропущенные буквы, пропущенные пробелы между словами): вывдится, минерадизация, фнкций, датируюся.
  - Сленг: оживляж, фолловер, балдёжный, заява.
  - «Нормальные» слова: никсы, дизрафия, складник, оцифровать.
  - Были проверены неизвестные слова корпуса текстов – **16 383** лемм найдены в Викисловаре.

Добавление дополнительных связей между словоформами



Добавление дополнительных связей между словоформами



Влияние многозначности на морфологический анализ

Пример получения морфологических характеристик для слов предложения:

Холодная капля стекла

холодная

- имя прилагательное, морфологические характеристики – ж. р., ед. ч. и др.

капля

- имя существительное, морфологические характеристики – ж. р., ед. ч., им. падеж и д.р.

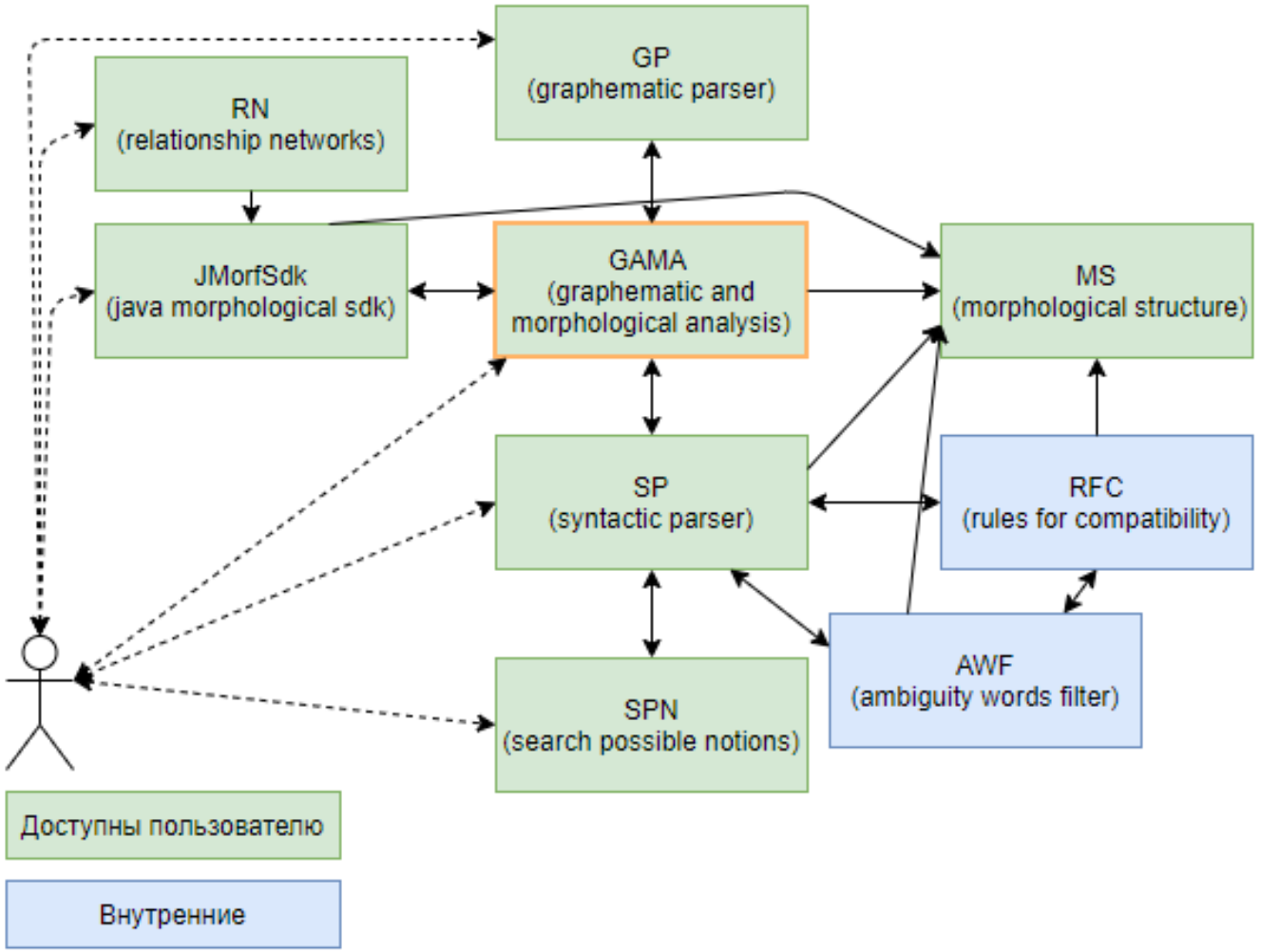
стекла

- имя существительное, морфологические характеристики – ср. р., ед. ч., родит. падеж и д.р.
- глагол, морфологические характеристики – ж. р., ед. ч., прош. время и д.р.

Методы решения морфологической многозначности

- **Контекстное снятие омонимии:**
  - с использованием дополнительных данных:
    - необходимость наличия статистики совместной встречаемости слов в большом корпусе текстов;
    - необходимость выделения окружения слова;
    - значительное повышение качества морфологического анализа;
    - небольшое снижение скорости морфологического анализа.
  - без использование дополнительных данных:
    - необходимость анализа окружения слова;
    - значительное повышение качества морфологического анализа;
    - существенное снижение скорости морфологического анализа.
- **Бесконтекстное снятие омонимии:**
  - необходимость наличия статистики встречаемости слов в большом корпусе текстов;
  - качество морфологического анализа ниже, чем у контекстного метода;
  - небольшое снижение скорости морфологического анализа.

Разрешение морфологической многозначности в TAWT



Результаты внесения изменений для улучшения работы библиотеки

- Обновление версии словаря и добавление дополнительного словаря:
  - в новой версии словаря **391 778** лемм**, 5 140 59**5 словоформ;
  - в дополнительном словаре ожидается более **20** тыс. лемм;
- Добавление улучшений и исправлений:
  - поддержка “ё”/“е” – улучшение на **0,21%;**
  - расширение набора форматов данных – улучшение более чем на **2%;**
  - повышение точности определения начальной формы глаголов, которые составляют **10,93%** слов корпуса текстов;
- Показатели получения морфологических характеристик библиотеки JMorfSdk с без снятия омонимии – **70,2%**, со снятием омонимии – **90,8%**. Добавление функции разрешения морфологической омонимии позволит значительно увеличить точность получения морфологических характеристик.

Спасибо за внимание!

Исследование инструментов морфологического анализа текстов на русском языке для повышения точности алгоритмов обработки в библиотеке JMorfSdk

к.т.н., доцент каф. 319 **Полицына Е.В.**,
к.т.н., доцент каф. 319 **Полицын С.А.**,
аспирант, каф. 319 **Поречный А.С.**,
студент, каф. 319 **Рыкунов А.Н.**