

Опыт автоматизированной классификации контекстов при морфологической разметке корпуса

Анна Голуб, 11А

Научный руководитель:

Дмитрий Владимирович Грановский

Место практики:

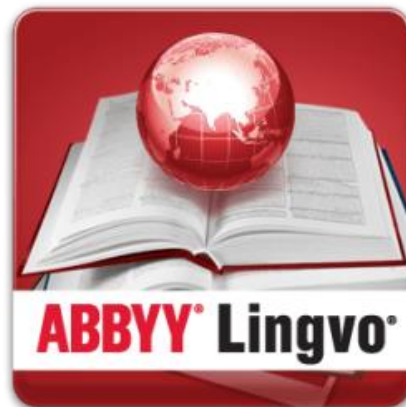
проект «Открытый корпус» (OpenCorpora)

Что такое корпус?

Корпус — база текстов.

Например:

- собрание текстов на одном языке и их переводов на другой язык



- собрание текстов на данном языке

OpenCorpora



Что такое разметка?

Разметка — это информация о лингвистических характеристиках различных составляющих текста.

фонетическая - звуки

морфологическая - слова

синтаксическая - предложения

Разметка OpenCorpora

- **морфологическая**
- размечаемая единица — слово
- разметка каждого слова — одна/несколько **интерпретаций**



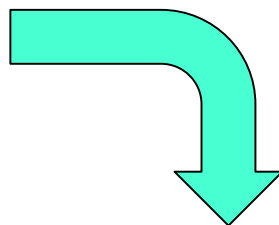
часть речи

The diagram consists of a central red word 'интерпретаций' from the list above. Two black lines extend from it: one to a small blue box on the left containing the text 'часть речи', and another to a larger blue box on the right containing the text 'набор значений обязательных для данной части речи грамматических категорий'.

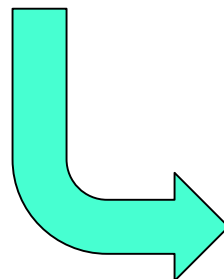
набор значений
обязательных для
данной части речи
грамматических
категорий

Создание разметки

**автоматическая
разметка**

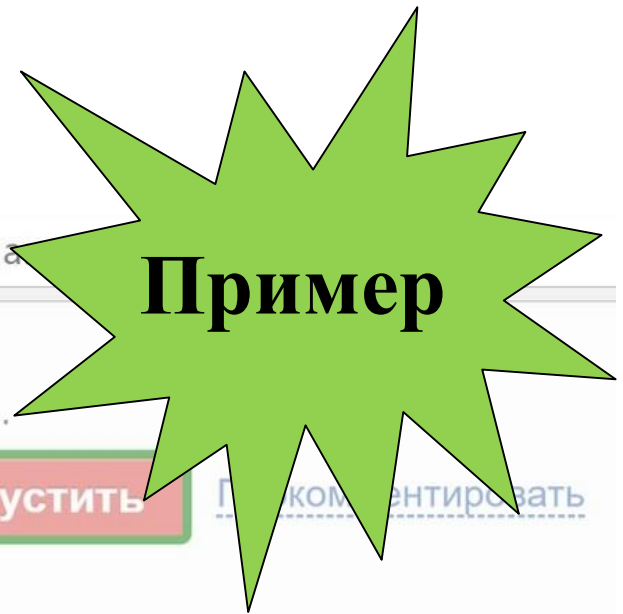


**краудсорсинг
(сайт OpenCorpora)**



модератор

Создание разметки



← → ↻ ⓘ Не защищено | www.opencorpora.org/tasks.php?act=a

... имеет собственные программы - **клиенты** для звонков .

именительный винительный Другое Пропустить [Прокомментировать](#)

... не способны ни на **жертвы** , ни на пожертвования ...

именительный **винительный** Другое Пропустить [Прокомментировать](#)

В Москве же Толины **дела** еще ухудшились , ибо ...

именительный винительный Другое Пропустить [Прокомментировать](#)

Создание разметки

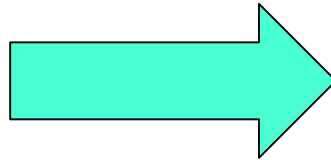


<...> «Рождественскую песню в
прозе: святочный **рассказ** с
привидениями»

Постановка задачи

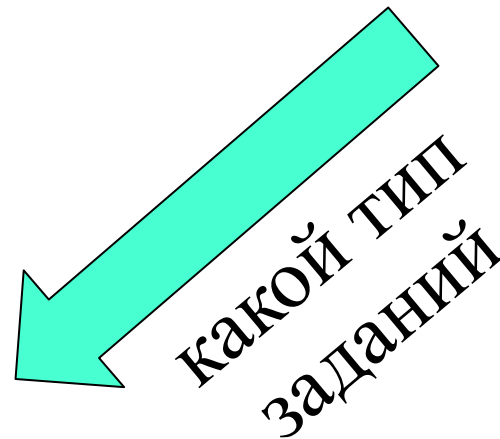
Фильтрация
части заданий,
попадающих к
модератору.

путь
решения



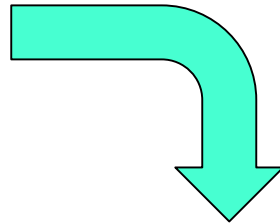
Выявление
критериев контекста,
дающих приемлемую
вероятность неверного
ответа.

«именительный падеж»
VS
«винительный падеж»
для **существительных**



Методика

Оценить частоту случаев, где все размечавшие ошиблись в выборе ответа.



Разработать набор **критериев контекста**, по которым можно выделять такие случаи.

Результат

- Была рассмотрена выборка из 35199 примеров.
- Случаев, где все ответы размечавших – неверные, 0.2%. Это больше, чем пороговое значение 0.1%.
- Имеет смысл дальнейшее исследование.

Оценить частоту случаев, где все размечавшие ошиблись в выборе ответа.

Результат

Были разработаны

признаки контекста

Разработать набор
критериев контекста,
по которым можно
выделять
ошибкоопасные случаи.

характеристики
размечаемого
слова

характеристики
контекста

характеристики
предыдущего
слова

Результат

Разработать набор
критериев контекста,
по которым можно
выделять
ошибкоопасные случаи.

- Написан код, проверяющий выполнение признаков у каждого предложения.
- Результат его работы - файл, где в каждой строке записано предложение и бинарные коды признаков в определенном порядке.

Результат



*Надо признать , обращение к современности
выбило из рук писателя самый главный его
козырь – [[знание]] скрупулезно описываемых
реалий прошлого . 2 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 1 0*

Результат

Разработать набор
критериев контекста,
по которым можно
выделять
ошибкоопасные случаи.

- **Критерий контекста** -
сочетание одного, двух или
трех признаков.
- Выявлены критерии
наиболее ошибкоопасных
контекстов.

Результат

**слово -
аббревиатура**

ФАС уже
возбудила
дело.

В США **сериал**
шёл с большим
успехом.

**перед словом -
аббревиатура**

**слово
- в кавычках**

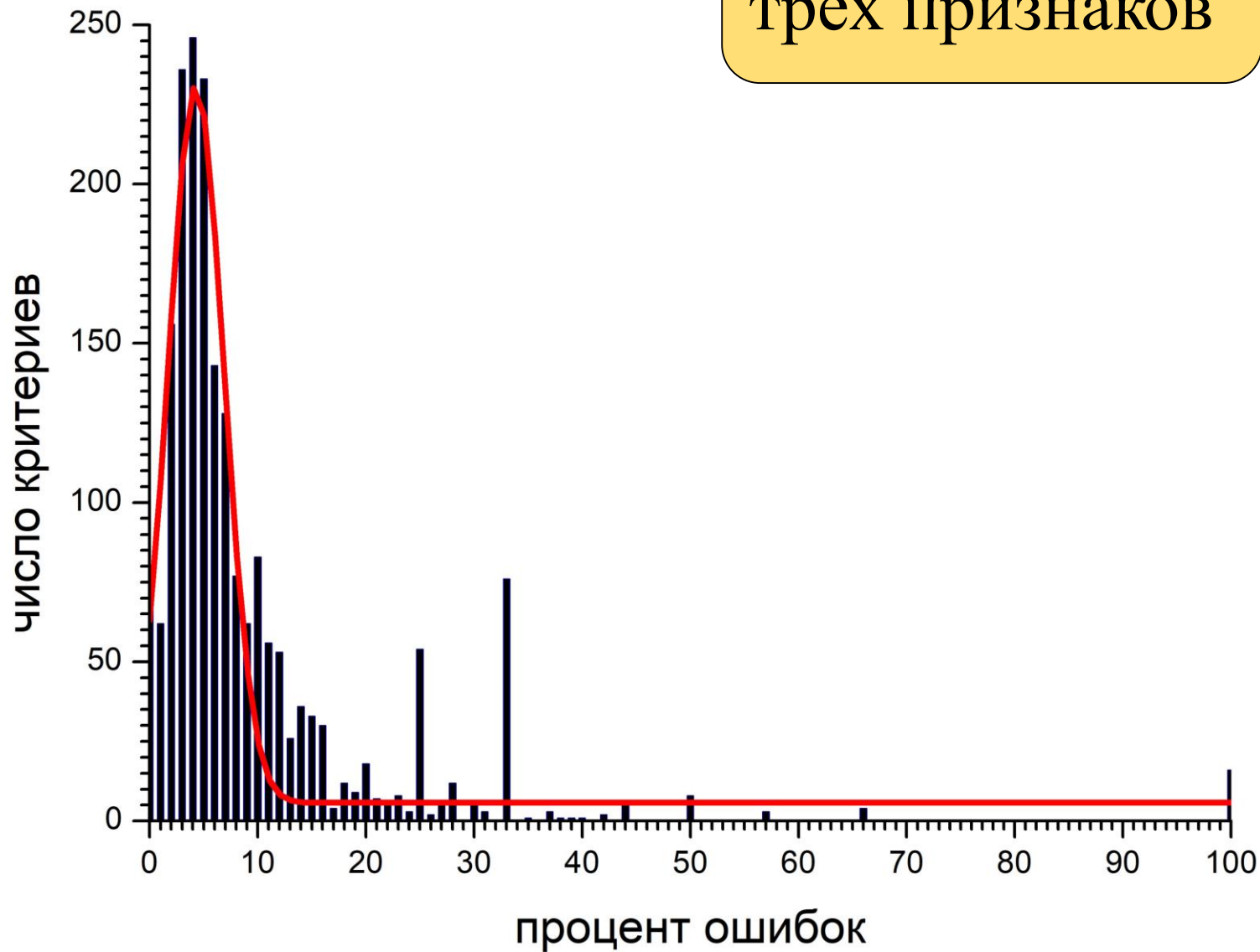
**«Ночь в
музее»**
сдвинули.

**перед словом
- число**

В итоге **«Зенит»**
набрал **61 очко**.

Результат

критерии из
трех признаков



Выводы

- Установлено, что существуют такие критерии контекста для заданий вида *"именительный надеж"* vs *"винительный надеж"* для существительных, что процент ошибок при выполнении соответствующего задания достаточно высок.
- Выявлены эти критерии.
- В дальнейшем эта информация может быть использована для автоматизации разметки *OpenCorpora*.



**Спасибо
за внимание!**