Московский Авиационный Институт
(Национальный исследовательский институт)

«Информатика: проблемы, методы, технологии» (IPMT-2021)

КЛАССИФИКАЦИЯ СУЩЕСТВИТЕЛЬНЫХ ИЗ ТЕКСТОВ МЕТОДАМИ МАШИННОГО ОБУЧЕНИЯ НА ОСНОВЕ ПРИЗНАКОВ КОНТЕКСТУАЛЬНОЙ СИНОНИМИИ

магистрант, каф. 319 **Милованова Е.Е.**

Естественный язык

Естественный язык – это язык, используемый для общения людей и не созданный целенаправленно.

- Обладает большим количеством семантических взаимосвязей;
- Позволяет описать объекты разными словесными конструкциями;
- Часто основывается не на точном значении единицы текста, а на её эмоциональной и стилистической окраске.

Проблема многозначности слов

- **Естественный язык** – сложная система, изначально не отличающаяся однозначностью толкования.
- **Неоднозначность** проявляется на всех уровнях анализа языка, что затрудняет автоматизированный анализ текстов.
- **Значение единицы языка** часто зависит от контекста употребления, слова в различных контекстах могут приобретать различные значения.
- **Выразительные средства**, использование синонимов, антонимов и т.д. для уточнения или усиления смысла ещё более усложняют автоматический анализ текста.

Контекстные синонимы

Синонимы – слова, равнозначные или похожие по значению.

Контекстный или контекстуальный синоним – это слово, которое равнозначно другому слову лишь в определенной ситуации, в конкретном контексте.

Контекст – это относительно законченная по смыслу часть текста или высказывания.

Синоним ≠ Контекстный синоним

Пример работы существующих программных средств по выделению контекстных синонимов

Исходное предложение	Онлайн – синонимайзер	Синонимайзер текстов
Вдруг он поднял голову, глаза его засверкали, он топнул ногою, оттолкнул секретаря с такою силою, что тот упал, и, схватив чернильницу, пустил ею в заседателя. Все пришли в ужас.	**Внезапно** он поднял голову, **очи** его засверкали, он топнул ногою, оттолкнул секретаря с такою силою, **собственно что что свалился**, и, схватив чернильницу, пустил ею в заседателя. Все **приехали** в **кошмар**.	**Внезапно** он поднял **сразу темя**, глаза его засверкали, он топнул ногою, оттолкнул секретаря с такою силою, что тот **свалился**, и, схватив чернильницу, пустил ею в заседателя. Все пришли в **страх**.

А.С. Пушкин «Дубровский»

Проблемы проектирования программных средств для выделения контекстуальных синонимов

- Зависимость от **объёмов исходных данных**. Как следствие, необходимость наличия и обработки больших и сверхбольших размеченных корпусов текстов, онтологий и тезаурусов;
- Проблема **подбора метрики** оценки корректности используемых эмпирических критериев, так как она зависит от результатов, представляемых человеком-экспертом и/или программного средства.

Признаки контекстуальных синонимов

- Полное или частичное совпадение характеристик, присущих словам-синонимам. То есть, контекстные синонимы должны совпадать по части речи, роду, числу, одушевлённости и т.д. Допустимы отклонения, например, для именованных сущностей или устойчивых выражений.
- Семантическая взаимозаменяемость с описываемым понятием в заданном контексте.
- Тождественность (не обязательно полная) толкования смысла слова с описываемым понятием.
- Принадлежность к одной предметной области.
- Стилистическое соответствие слова контексту.
- Поведенческая схожесть для соблюдения правильности построения предложения.

Описание классов сущностей в тексте

- **Person** – одушевлённый персонаж, выполняющий активную роль в повествовании.
- **Object** – неодушевлённый предмет, имеющий важную роль в тексте или над которым производятся какие-либо действия.
- **Something** – объекты или одушевлённые лица, не играющие важной роли в тексте.

Средства разработки

TAWT

- Язык разработки *Java;*
- Фреймворк *TAWT* – библиотека для языка *Java*, реализующая графематический, морфологический и семантико-синтаксические анализы текста на русском языке.
- Weka (Waikato Environment for Knowledge Analysis) — свободное программное обеспечение для анализа данных и машинного обучения, написанное на Java в Университете Уаикато (Новая Зеландия), распространяющееся по лицензии GNU GPL.
- *Spring Framework* – универсальный фреймворк с открытым исходным кодом для *Java*-платформы.
- Среда разработки: *IntelliJ IDEA.*

Вектор признаков

- именованность;
- род;
- одушевленность;
- общая частота употребления;
- количество раз, когда слово было главным в предложении;
- общее количество зависимых от слова глаголов;
- общее количество раз, когда слово зависимо от глагола;
- общее количество зависимых от слова прилагательных;
- общее количество раз, когда слово зависимо от прилагательного;
- общее количество зависимых от слова существительных;
- общее количество раз, когда слово зависимо от существительного.

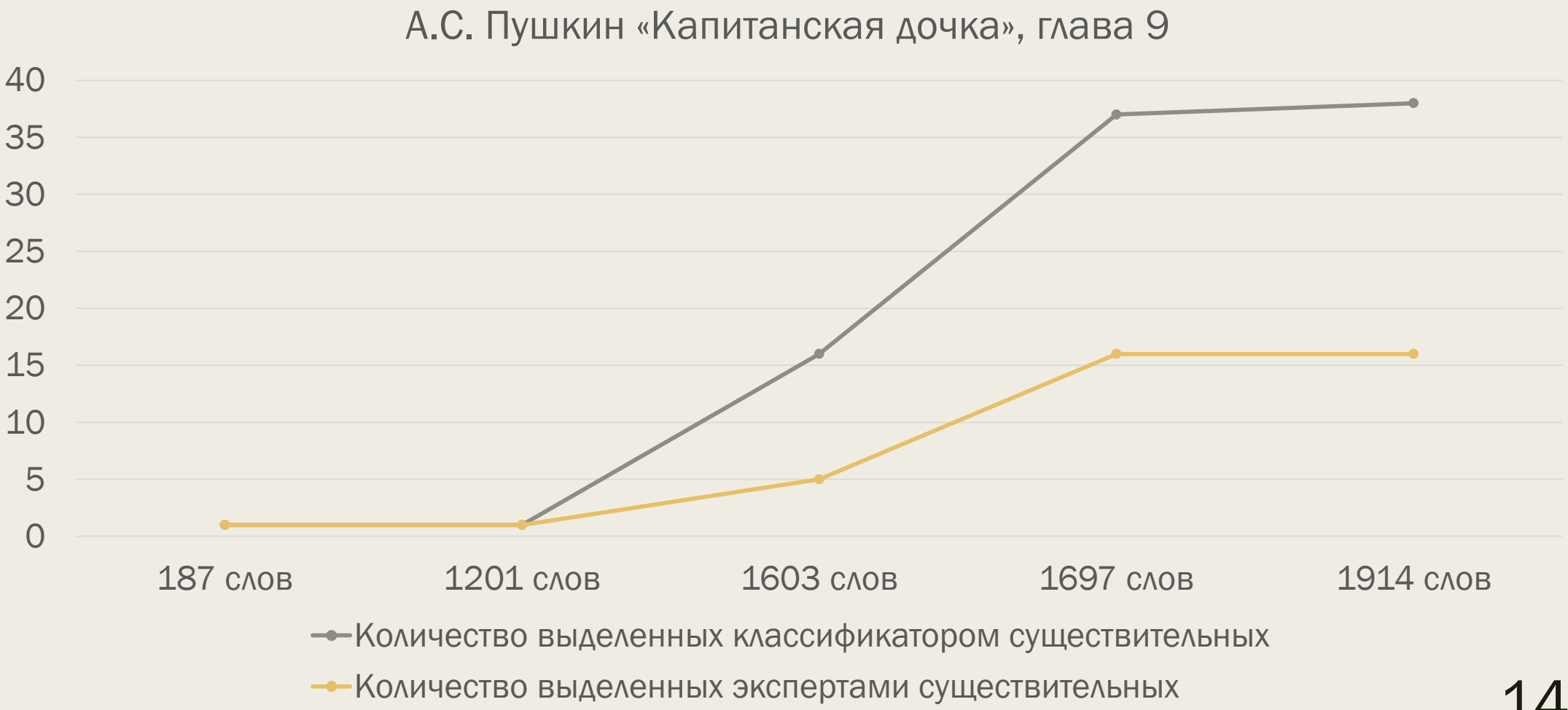
Алгоритм векторизации существительных из текстов на естественном языке

Описание используемых для обучения классификатора текстов

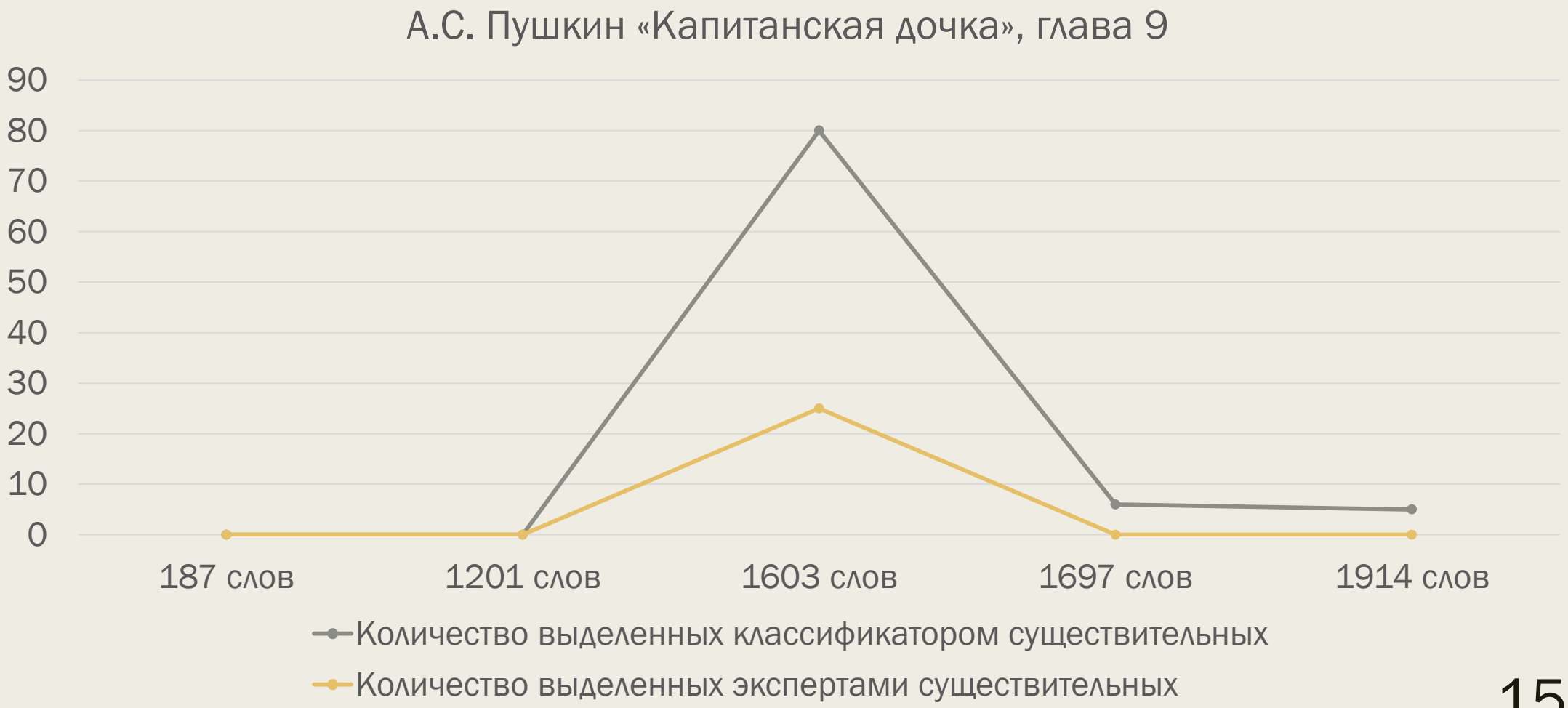
Автор	Название произведения	Количество существительных в тексте	Общее количество слов в тексте
Н.Н. Носов	«Живая шляпа»	34	187
Л.А. Чарская	«Лидианка»	391	1201
Л.А. Чарская	«Поповна»	505	1603
Л.А. Чарская	«Нелюбимая»	559	1692
А.П. Чехов	«Каштанка»	519	1914

Алгоритм классификации существительных из текстов на естественном языке

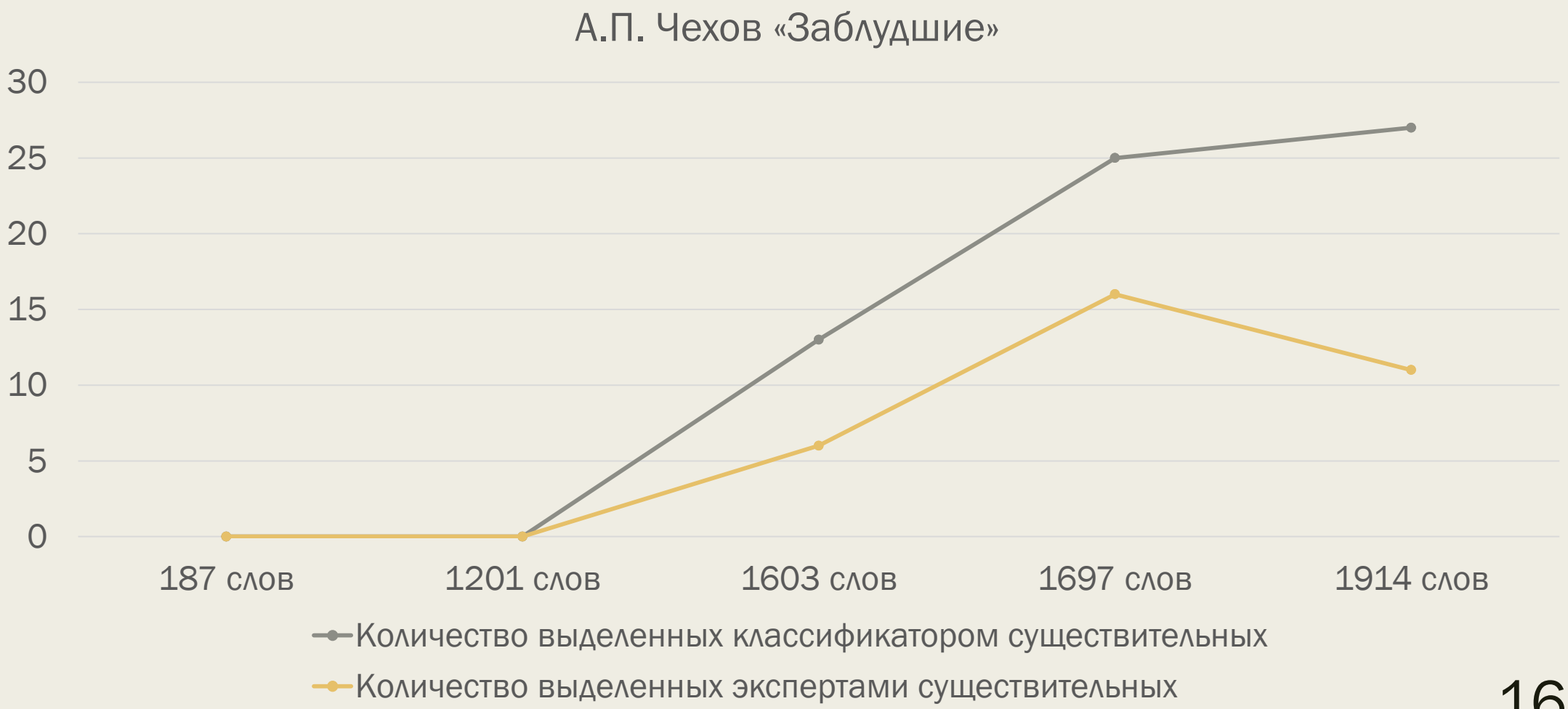
Оценка точности работы моделей классификатора на основе текстов художественной литературы (класс «person»)



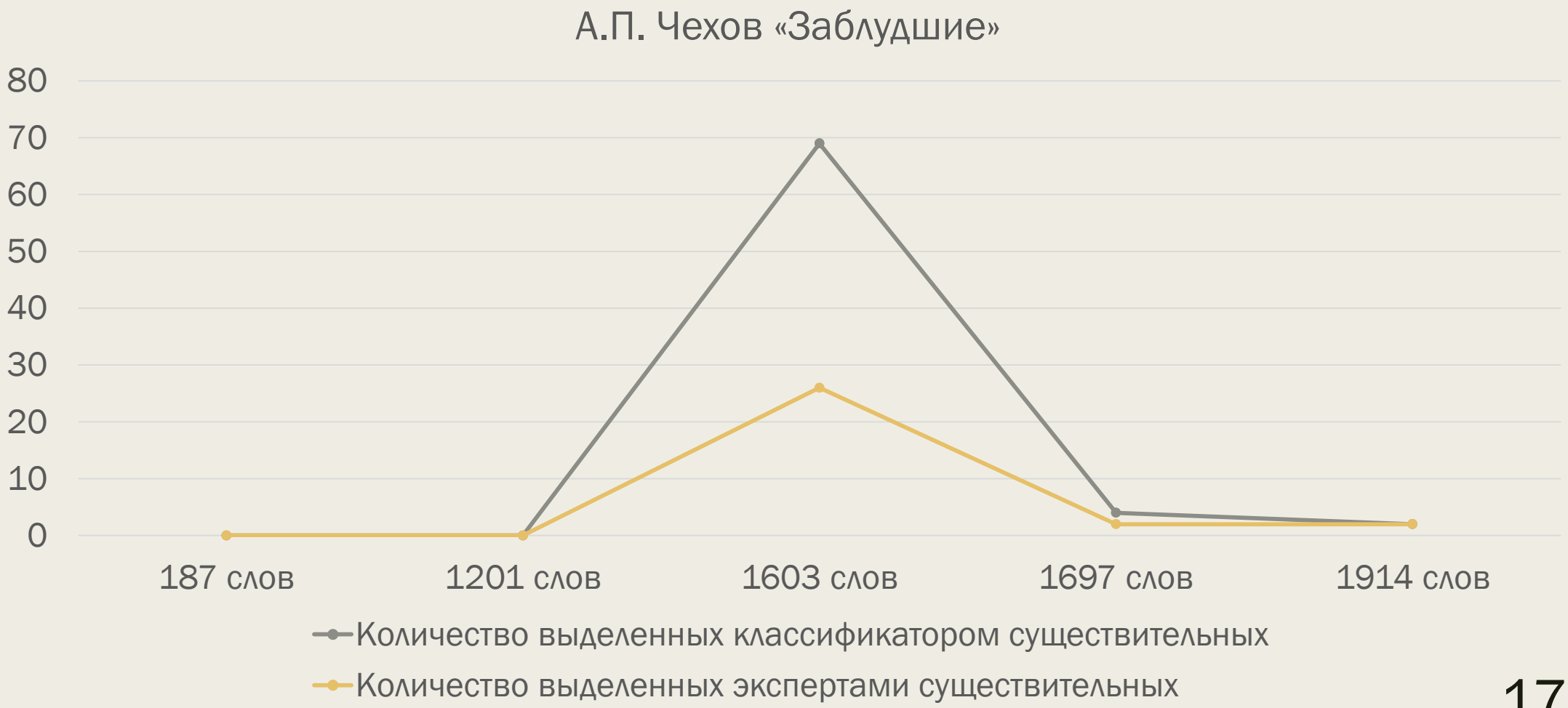
Оценка точности работы моделей классификатора на основе текстов художественной литературы (класс «object»)



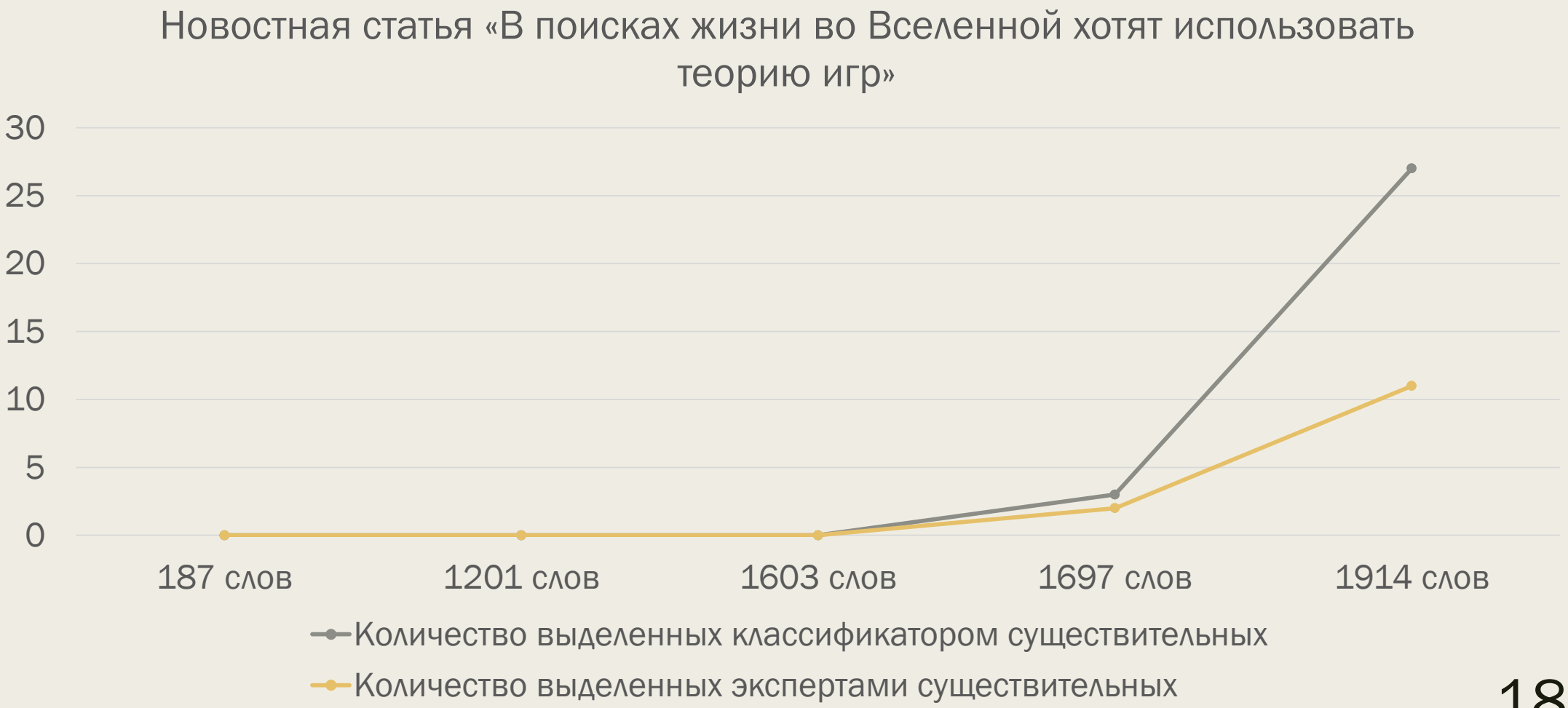
Оценка точности работы моделей классификатора на основе текстов художественной литературы (класс «person»)



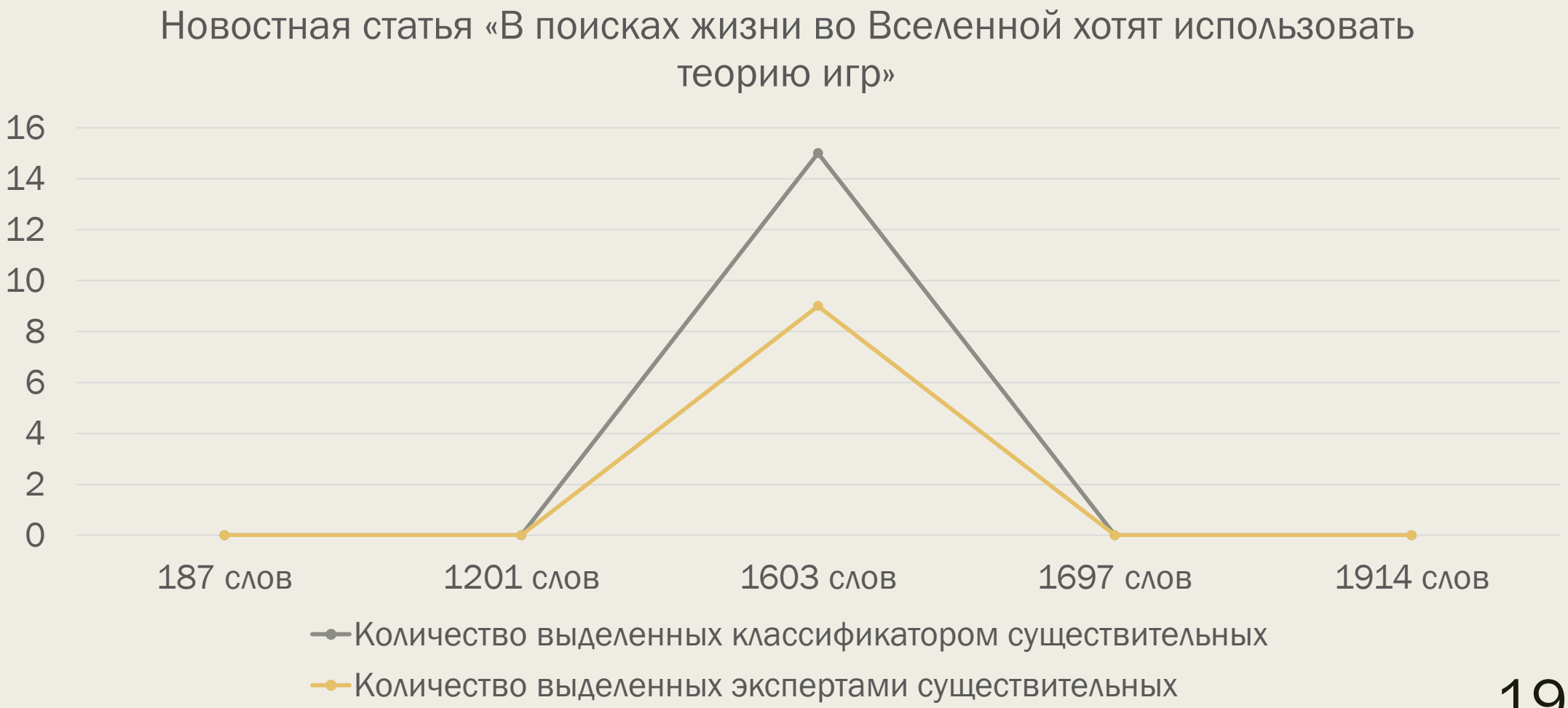
Оценка точности работы моделей классификатора на основе текстов художественной литературы (класс «object»)



Оценка точности работы моделей классификатора на основе текстов художественной литературы (класс «person»)



Оценка точности работы моделей классификатора на основе текстов художественной литературы (класс «object»)



Пути развития

- Расширение библиотеки обученных моделей классификации на текстах разных жанров.
- Проведение дополнительного лингвистического анализа для совершенствования списка признаков для векторизации.
- Развитие используемых методов векторизации текстов.
- Создание ансамбля из обученных моделей классификации для качественного улучшения результатов.

Выводы

1. Сформирован **список характеристик**, включающий в себя морфологические и синтаксические признаки, которые могут быть использованы для формирования наборов данных для обучения классификатора.
2. Предложен **алгоритм векторизации** слов на основе признаков контекстуальной синонимии для формирования набора данных, предназначенных для обучения классификатора.
3. Проведено исследование **точности** результатов работы моделей классификатора, обученных на текстах художественной литературы.
4. Отмечена **зависимость** результатов классификации от размеров текстов, являющихся основой для обучения классификатора, и текстов, для которых выполняется процедура классификации существительных.
5. Выявлено, что для **повышения качества** выделения разных классов требуется использование моделей, обученных на разных текстах.

Спасибо за внимание!

«Классификация существительных из текстов методами машинного обучения на основе признаков контекстуальной синонимии»

магистрант, каф. 319 **Милованова Е.Е.**