ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное учреждение высшего образования

Национальный исследовательский университет «Высшая школа экономики»

Факультет гуманитарных наук

Образовательная программа «Фундаментальная и компьютерная лингвистика»

Сухарева Мария Игоревна

РАСПОЗНАВАНИЕ ИМЕНОВАННЫХ СУЩНОСТЕЙ В СРЕДНЕВЕКОВЫХ НЕМЕЦКИХ ТЕКСТАХ

Выпускная квалификационная работа студента 4 курса бакалавриата группы БКЛ-201

Академический руководитель образовательной программы канд. филологических наук, доц. Ю.А. Ландер

_____

« » _____ 2024 г.

Научный руководитель

к.фил.н., доцент К.К. Кашлева

_____

Москва 2024

Содержание

Аннотация

Модели распознавания именованных сущностей (NER), хотя и показывают хорошие результаты на современных корпусах, часто сталкиваются с трудностями при работе с историческими данными, в первую очередь из-за нехватки аннотированных данных для обучения. В данном исследовании представлен новый NER датасет, основанный на корпусе текстов на средневерхненемецком языке. Этот датасет содержит 387 лемм, разделенных на 6 тегов: PERSON, REGION, CITY, WATER, GROUP и PROPERTY. Кроме того, в данной работе используется датасет (Besnier, Mattingly 2021), основанный на “Песни о Нибелунгах”, для обучения трех NER моделей: Distilroberta-base-mhg-charter-mlm, GHisBERT и SequenceTagger от Flair. Эти модели были протестированы на собранном датасете, основанном на текстах о Дитрихе Бернском. Модель Flair, единственная из представленных, которая изначально была обучена на современном немецком языке, демонстрирует лучшие результаты с F1-мерой, равной 0,74. Однако результаты показали, что все три модели не смогли выучить тег GROUP из-за недостаточного количества примеров в обучающих данных. Собранный в данной работе датасет предоставит больше данных для обучения моделей на средневерхненемецком языке, что проложит путь к более масштабным компьютерным исследования средневековой немецкой литературы.

1. Введение

Одно из центральных мест в обработке естественного языка (Natural Language Processing, NLP) занимает такая задача, как распознавание именованных сущностей (Named Entity Recognition, NER). NER применяется для автоматизации более сложных NLP задач, например, таких, как интеллектуальный анализ текста, классификация текстов, поиск информации и других задач машинного обучения. Распознавание именованных сущностей обеспечивает систематический подход к определению ключевых элементов в тексте, таких как имена людей, места, бренды и даже конкретные денежные единицы. Благодаря тому, что NER модели облегчают

извлечение ключевых данных из корпуса текстов, они оказывают важную помощь в структурировании неупорядоченных данных и выделении важной информации.

Исторически сложилось так, что исследования в сфере NER в основном проводились на современных языках. По большей части это связано с тем, что модели, обученные на современных языках, открывают возможность для быстрого создания коммерческих продуктов. В то время как результаты исследований, основанных на средневековых данных, скорее относятся к теоретической области. Тем не менее, извлечение информации из средневековых произведений может помочь исследователям получить более глубокое представление о данных текстах. Например, построение графов персонажей позволяет визуализировать их взаимодействия в текстах, тем самым улучшая понимание динамики персонажей (Kenna, MacCarron 2017). Более того, NER модели также могут помочь в создании индексированной структуры исторических текстов, например, усовершенствовать методы поиска информации общего назначения. Ярким примером такого корпуса средневерхненемецкого языка, в котором отдельно используются аннотации для именованных сущностей, является Middle High German Conceptual Database (Zeppezauer-Wachauer 2024).

Кроме того, одно из постоянных препятствий для применения машинного обучения при исследовании средневековых языков связано с доступностью данных. Однако в последние годы возрос интерес к компьютерным исследованиям в области средневековых языков, в том числе к средневерхненемецкому языку, и средневековым языкам в компьютерной сфере стало уделяться немного больше внимания. Об адаптации компьютерных методов к историческим текстам подробно рассказывается в исследовании (Piotrowski 2012), а в более недавней статье (Novotný et al. 2023) обращается внимание на потенциальные сложности при компьютерной обработке исторических текстов. При этом появление усовершенствованных технологий, в частности оптического распознавания символов (Optical Character Recognition – OCR), увеличило количество доступных текстов. Тем не менее потребность в аннотированных текстах по-прежнему остается высокой, что обусловливает актуальность исследования.

Цель данной работы заключается в том, чтобы создать аннотированный датасет, который будет полезен при обучении NER моделей, специально предназначенных для средневерхненемецкого языка. Новый датасет будет включать в себя именованные сущности из средневерхненемецких текстов о Дитрихе Бернском. Кроме того, планируется обучить три NER модели (Distilroberta-base-mhg-charter- mlm, GHisBERT, Flair) на именованных сущностях из датасета (Besnier, Mattingly 2021), основанного на тексте “Песни о Нибелунгах”, и протестировать их на текстах о Дитрихе с помощью созданного датасета. Для достижения цели были решены следующие задачи: 1) собрать корпус текстов о Дитрихе Бернском; 2) составить датасет именованных сущностей и разметить его; 3) обучить три NER модели на датасете (Besnier, Mattingly 2021); 4) посчитать метрики качества на собранном датасете.

Данное исследование представляет собой комплексную работу, направленную на увеличение объема аннотированных средневековых данных и дальнейшее развитие области распознавания именованных сущностей в средневековых текстах.

2. Теоретическая часть

2.1. Обзор литературы

К составлению аннотированных датасетов для распознавания именованных сущностей в области средневековых языков исследователи подходят по-разному. В статье (Besnier, Mattingly 2021) представляется вручную собранный датасет с именованными сущностями на трех языках – средневековой латыни, средневерхненемецком и древнескандинавском. Датасет, содержащий имена людей (PERSON), названия мест (PLACE) и именования групп людей (GROUP), изначально был создан для извлечения персонажей из трех связанных средневековых текстов – *Decem Libri Historiarum*, *Völsunga saga* и *Nibelungenlied*. Сборка и разметка данных вручную, хотя и требует больших затрат времени, позволяет получить более высокий уровень точности. Тщательное извлечение сущностей человеком обеспечивает меньшую вероятность пропустить и большую вероятность правильно классифицировать эти сущности, в то время как автоматические компьютерные

методы не всегда обеспечивают такую точность. Авторы считают, что такой датасет может быть полезен для создания автоматических инструментов распознавания именованных сущностей для средневековых языков с небольшим объемом аннотированных данных.

С другой стороны в работе (Novotný et al. 2023) показывается, что можно автоматически создать аннотированный NER корпус с помощью методов извлечения информации, используя неаннотированный корпус исторических текстов и списки известных исторических личностей и местностей. Авторы разрабатывают новый NER корпус с тегами PERSON и LOCATION, состоящий из 3,6 миллионов предложений из хартий позднего средневековья, написанных в основном на древнечешском, латыни и средневерхненемецком языках. Однако автоматизированный подход, несмотря на то, что он значительно быстрее и способен обрабатывать большие объемы данных, не является совершенным. Такой подход для средневековых текстов, в которых необходимо учитывать множество нюансов, требует последующей ручной проверки. В этом исследовании была обучена NER модель с F1-score = 80-93% при проверке на вручную размеченных данных. Кроме того, авторы показывают, что использование взвешенной функции потерь помогает бороться с классовым дисбалансом в задачах классификации токенов.

Более смешанный подход используется в исследовании (Schulz, Ketschik 2019), в котором авторы создают теггер частей речи (Part Of Speech tagger, POS-tagger) для средневерхненемецкого языка. Они обсуждают различные аспекты работы с средневековыми языками, такие как количество данных, необходимое для обучения, и влияние качества данных на производительность теггера. Для обучения модели авторы используют вручную аннотированные данные, что обеспечивает хорошее качество, но кроме этого, они обсуждают, как можно использовать неаннотированные данные в качестве дополнительных данных для обучения для повышения эффективности теггера в конкретной области. Такая стратегия не только увеличивает количество данных, доступных для обучения, но и знакомит модель с более широким набором примеров, относящихся к конкретной предметной области. Полученная в результате POS модель обеспечивает точность разметки около 91% на различных

тестовых наборах, в которые включены различные жанры, периоды времени и разные виды средневерхненемецкого языка.

В следующих работах используются вручную размеченные данные, но основное внимание уделяется применяемым моделям. В статье (Aguilar 2022) обсуждается использование компьютерных методов для обработки средневековых текстов на латыни, среднефранцузском и средневековом испанском языках. В исследовании предлагается использовать методы машинного обучения, в частности контекстуальные и статические эмбеддинги в паре с классификатором Bi-LSTM-CRF, а также обучение моделей BERT и RoBERTa, для автоматического распознавания именованных сущностей (PERSON и LOCATION) в исторических хартиях. Исследование проводилось на основе корпуса из около 2,3 миллионов токенов, собранных из пяти сборников хартий, охватывающих X-XV века. Результаты исследования показывают высокую эффективность мультиязычных классификаторов, как основанных на моделях общего назначения, так и созданных специально, без заметного снижения производительности по сравнению с их одноязычными аналогами.

В статье (Aguilar, Stutzmann 2021) представлен процесс создания аннотированного корпуса с тегами PERSON и LOCATION и описано применение компьютерных методов для автоматического извлечения именованных сущностей из средневековых хартий на среднефранцузском языке. Исследование было основано на корпусе, включающем около 500 тысяч токенов из 1200 хартий, собранных из трех сборников хартий XIII и XIV веков. В ходе исследования были обучены три модели – SpaCy, Flair и Bi-LSTM-CRF. Оценка показала, что все три модели демонстрируют высокую производительность на данных, которые модели не видели при обучении.

2.2. Средневековая немецкая литература

В (Gibbs, Johnson 1997) рассматриваются произведения разных жанров немецкой литературы, начиная с раннего Средневековья и заканчивая поздним Ренессансом. В разделе “Middle High German Literature”, который посвящен литературе XII-XIII веков, описываются такие жанры, как рыцарский роман,

героический эпос и лирическая поэзия. Жанру героического эпоса посвящена также работа (McConnell 2002).посвящена.

Среди эпической литературы на средневерхненемецком особо выделяется “Песнь о Нибелунгах” и цикл текстов о Дитрихе Бернском.

“Песни о Нибелунгах” посвящено множество исследовательских работ, рассматривающих ее с разных сторон. Gibbs, Johnson (1997) уделяют особое внимание образам главных героев эпоса “Песнь о Нибелунгах” и их роли в развитии сюжета. Авторы рассматривают главных героев – Кримхильду, Зигфрида, Хагена и Гюнтера, но также и некоторых второстепенных персонажей – Этцеля и Брюнхильду. В книге исследуются мотивы их поступков и влияние на развитие событий. Например, Кримхильда мстит за убийство своего мужа Зигфрида, несмотря на предостережения других персонажей, что приводит к гибели многих героев.

Похожие рассуждения можно найти в работе (Classen 2009), которая посвящена тому, как в “Песни о Нибелунгах” сочетаются мифические и исторические элементы. В статье описываются различные ссылки в тексте на реальные случаи, с какими историческими событиями мог бы быть знаком анонимный поэт. Автор отмечает, что хотя поэма основана на древних германских мифах, она также содержит элементы реальной истории. Например, в ней упоминаются реальные исторические личности, такие как Аттила и Теодорих Великий, а также борьба между бургундами и гуннами.

В своей основополагающей работе (McConnell et al. 2001) авторы подробно проанализировали мифы о нибелунгах по целому ряду тем, включая литературные и внелитературные ссылки, персонажей и географические названия, а также значимые мотивы и концепции, исторический фон и культурное восприятие на протяжении веков. В книге рассказывается о происхождении нибелунгских мифов и их развитии в разных культурах. Их возникновение связывают с древними германскими легендами и сказаниями. Эти мифы передавались устно из поколения в поколение, а затем были записаны в средневековых эпических поэмах, таких как “Песнь о Нибелунгах”. Происхождение легенд о нибелунгах связывают с такими историческими событиями, как миграция народов, войны и завоевания. Также в этой книге подробно исследуются мотивы поступков персонажей “Песни о Нибелунгах”, их отношения

друг с другом и влияние на развитие сюжета. Авторы анализируют различные интерпретации этих образов в искусстве и культуре разных эпох.

Помимо общего анализа эпических текстов, исследователи рассматривают и отдельные языковые и культурные детали. Например, в (Heinrichs 1955) анализируются лингвистические особенности, связанные с именами героев ряда древних саг. Автор сосредотачивается на анализе фонетических трансформаций имен и связывает их с возможными историческими событиями. В частности, исследуется эволюция имен *Sigiward, Sigfrid* и *Sivrit*, *Gernot, Kriemhilt* и *Grimhild*, как такие имена, как *Godofrid* превратились в *Godevert* и далее в *Govert*. В тексте представлены доказательства процессов метатезиса (изменения порядка фонем в слове), которые могли повлиять на преобразования этих имен. Автор также обсуждает влияние региональных диалектов на соответствующие фонетические изменения. Он приходит к выводу, что на трансформацию имен повлияли не только диалектные, но и социокультурные факторы, такие, как желание “очистить” язык от просторечий.

В другой работе (Schramm 1965) рассматривается возможная связь между именами *Kriemhilt* и *Ildico,* и предполагается, что *Ildico* – это случайный или неофициальный вариант имени. Но на основе этимологического исследования и некоторых исторических источников автор делает вывод, что *Ildico* уже была известна, а связь *Kriemhilt* с *Ildico* уже появилась позже. Кроме того, рассматривается связь *Kriemhilt* с *Grimhild*. Предполагается, что корни этих имен могли иметь одинаковое происхождение или могли быть изменены носителями языка, однако в статье аргументы в пользу таких гипотез обсуждаются с осторожностью, указывая на то, что, хотя некоторые лингвистические и исторические доказательства связи этих имен существуют, все же точных выводов сделать нельзя.

Кримхильде в целом часто посвящаются отдельные статьи, так как она является одним из главных героев эпоса “Песнь о Нибелунгах”. В статье (Layher 2009) рассматривается ее образ в контексте средневековой культуры. Она известна своей решительностью, гордостью и мстительностью. В статье анализируется, как этого персонажа использовали в качестве примера в проповедях XIII века для иллюстрации

определенных моральных принципов. Также автор рассматривает различные интерпретации образа Кримхильды и то, как они менялись со временем.

Основные темы, которые раскрываются в “Песни о Нибелунгах”, это месть, любовь и верность, власть, судьба и героические подвиги. Все эти мотивы, а также некоторые персонажи, связывают этот текст с циклом текстов про Дитриха Бернского.

Если “Песнь о Нибелунгах”, в первую очередь, история о любви и предательстве, то в цикле текстов про Дитриха любовные мотивы менее выражены, больше обращается внимание на героические подвиги и сражения. Считается, что Дитрих Бернский является отражением короля остготов Теодориха Великого (McConnell 2002). Некоторые из этих эпосов, в которых Дитрих изображен в качестве главного героя, основаны на стремлении Теодориха создать империю на севере Италии. А большинство различий между известным жизнеописанием Теодориха и образом Дитриха в сохранившихся легендах объясняются давней устной традицией, которая продолжалась вплоть до XVI века.

И (Gibbs, Johnson (1997), и (McConnell 2002) описывают 11 текстов про Дитриха, разделяя их на две части: исторические и мифические. К историческим текстам относятся *Dietrichs Flucht*, *Rabenschlacht*, *Alpharts Tod* и отрывочно сохранившаяся *Dietrich und Wenezlan*. Все четыре относятся к периоду после появления Дитриха в “Песни о Нибелунгах”. Они называются историческими, потому что в них рассказывается о войне, а не о приключениях, и считается, что они содержат искаженную версию жизни короля остготов Теодориха Великого. Большинство сохранившихся рассказов о Дитрихе носят фантастический характер и включают сражения с мифическими существами и другими героями. К мифическим текстам относятся: *Goldemar*, *Eckenlied*, *Sigenot*, *Virginal*, *Laurin und Walberan*, *Rosengarten* и *Wunderer*.

В статье (Voorwinden 2007) рассматриваются эпосы *Dietrichs Flucht* и *Rabenschlacht*. Эти тексты считаются историческими из-за их вероятной связи с реальными событиями, такими как вторжение Теодориха в Италию, его три битвы с Одоакром и гибель одного из сыновей короля Этцеля в битве 454 года. Однако автор

предполагает, что эти эпосы также могли быть созданы как имитация историографии, учитывая некоторые исторические неточности.

Все эти произведения являются важными памятниками германской литературы и отражают основные ценности и представления средневекового общества о жизни.

2.3. Тексты для корпуса

Для корпуса было отобрано семь текстов. В него вошли *Nibelungenlied* (редакция B); четыре исторических эпоса о Дитрихе: *Dietrichs Flucht* (редакция R), *Rabenschlacht* (редакция R), *Alpharts Tod*, *Dietrich und Wenezlan*; и 2 мифических эпоса: *Rosengarten* (редакция A) и *Wunderer* (редакция Dresdener Heldenbuch). Все тексты связаны друг с другом одним персонажем – Дитрихом Бернским. Во всех эпосах он появляется как главный герой, кроме “Песни о Нибелунгах”, где он является второстепенным персонажем.

2.3.1. Nibelungenlied “Песнь о Нибелунгах”

Поэма разделена на две части. В первой части принц Зигфрид приезжает в Вормс, чтобы добиться руки бургундской принцессы Кримхильды у ее брата короля Гюнтера. Гюнтер соглашается позволить Зигфриду жениться на Кримхильде, если Зигфрид поможет Гюнтеру заполучить в жены королеву-воительницу Брюнхильду. Зигфрид делает это и женится на Кримхильде, однако Брюнхильда и Кримхильда становятся соперницами, что в конечном итоге приводит к убийству Зигфрида бургундским вассалом Хагеном при участии Гюнтера. Во второй части вдова Кримхильда выходит замуж за Этцеля, короля гуннов. Позже она приглашает своего брата и его двор посетить королевство Этцеля, намереваясь убить Хагена. Ее месть приводит к гибели всех бургундов, которые приезжали ко двору Этцеля, а также к разрушению королевства Этцеля и смерти самой Кримхильды.

Поэма была утеряна к концу XVI века, но рукописи датируемые еще XIII веком, были вновь обнаружены в течение XVIII века (Bryn 2008). Существует 37 известных рукописей “Песни о Нибелунгах” и ее вариантов, некоторые из них считаются завершенными. Самой старой версией является та, которая сохранилась в редакции B. Из трех древнейших рукописей каждая представляет собой отдельную редакцию:

A, B и C. Эта классификация основана на подписях к рукописям. По параметру формулировки последнего стиха в каждом источнике редакции обычно делят на две крупные группы: *Liet-Fassung* или *Nôt-Fassung* (Lachmann 1878). Редакция A значительно короче других, B и C считаются наиболее полными, они различаются последним стихом. В уже существующем датасете (Besnier, Mattingly 2021) использовалась редакция C, поэтому для нашего корпуса мы взяли редакцию B.

“Песнь о Нибелунгах” традиционно датируется примерно 1200 годом и, как и другие героические эпосы средневерхненемецкого языка, анонимна (Curschmann 1987). Хотя часто утверждается, что “Песнь о Нибелунгах” создал один автор, степень расхождения в тексте и его предыстория в устной традиции означает, что идеи авторского замысла следует применять с осторожностью (Müller 2009). Возможно также, что данную поэму писали несколько поэтов под руководством одного редактора.

Язык “Песни о Нибелунгах” характеризуется своей шаблонностью, характерной чертой устной поэзии, означающей, что похожие или идентичные слова, эпитеты, фразы и даже строки можно найти в различных местах на протяжении всей поэмы. Эти формулы могут быть гибко использованы в поэме для различных целей. Поскольку обычно считается, что “Песнь о Нибелунгах” была задумана как письменное произведение, эти элементы обычно воспринимаются как признаки “вымышленной устности” (*fingierte Mündlichkeit*), которые подчеркивают связь поэмы с ее традиционно устным сюжетом (Müller 2009).

Текст содержит около 2400 четырехстрочных строф и 39 авентюр. Нибелунгова строфа состоит из трех “длинных строк” (Langzeilen), которые состоят из трех метрических стоп, цезуры (паузы, делящей стих на части) и трех метрических стоп, следующих за цезурой. Четвертая строка добавляет дополнительную стопу, следующую за цезурой, делая ее длиннее трех остальных и обозначая конец строфы. Последнее слово перед цезурой обычно женского рода (за ударным слогом следует безударный), в то время как последнее слово строки – обычно мужское (ударный слог) (Millet 2008). Характер строфы создает структуру, в соответствии с которой повествование развивается блоками: первые три строки развивают историю, в то

время как четвертая представляет собой предзнаменование катастрофы в конце или комментарии к событиям. Переходы между строфами очень редки (Müller 2009).

2.3.2. Dietrichs Flucht "Бегство Дитриха"

Данная поэма описывает правление предков Дитриха в его королевстве на севере Италии, его предательство и изгнание злым дядей Эрменрихом и его бегство к гуннам, где его тепло принимают Этцель, король гуннов, и его жена Хельха. С помощью Этцеля Дитрих предпринимает две попытки отвоевать свое королевство у Эрменриха, но они заканчиваются неудачей, и каждый раз он вынужден возвращаться в изгнание к гуннам.

"Бегство Дитриха" дошло до нас вместе с *Rabenschlacht* в четырех полных рукописях и отдельно в одной фрагментарной рукописи. Основной редакцией считается рукопись R, поэтому она была выбрана для нашего корпуса. Происхождение самых ранних рукописей, а также диалект, на котором написана поэма, указывают на то, что она была написана в Австрии где-то до 1300 года (Heinzle 1999).

Текст содержит около 10 000 строк. Поэма необычна тем, что она написана рифмованными двустишиями, а не строфами, как большинство немецких героических эпосов. За единственным исключением в самом начале, рассказчик абсолютно уверен в правдивости своей истории, повторяя снова и снова, что то, что он рассказывает – правда. Его утверждения о том, что он ссылается как на письменные, так и на устные источники, могут убедить его аудиторию в том, что он рассказывает историческую правду, в то же время придавая поэме авторитет устной традиции (Lienert 2003). В то же время поэма стремится создать нечто вроде целостного рассказа о героическом мире, включая персонажей из *Wolfdietrich*, *Ortnit* и *Nibelungenlied*: Загфрид, Гюнтер и Геронт – все они фигурируют в поэме.

2.3.3. Rabenschlacht "Битва при Равенне"

"Битва при Равенне" начинается через год после окончания "Бегства Дитриха". Поэма повествует о неудачной попытке изгнанника Дитриха отвоевать свое королевство в Северной Италии у своего дяди Эрменриха с помощью армии,

предоставленной Этцелем. В ходе этой попытки младший брат Дитриха и малолетние сыновья Этцеля от его жены Хельхи были убиты бывшим вассалом Дитриха Витеге недалеко от Равенны. Витеге затем бежит в море и спасается благодаря русалке вместо того, чтобы сражаться с Дитрихом. Возможно, эта поэма является смутным отражением гибели сына Аттилы Эллака в битве при Недао в 454 году в сочетании с осадой Равенны Теодорихом Великим в 491-493 годах.

Большинство современных исследователей считают, что “Битва при Равенне” была написана раньше, чем “бегство Дитриха”. (Hoffmann 1974) предполагает, что “Битва при Равенне”, возможно, была написана около 1270 года, прежде чем была переработана и помещена в одну книгу вместе с “Бегством Дитриха” в 1280-ых годах. Хотя некоторые исследователи все же считают, что эти два эпоса были написаны одновременно, но возможно, существовали и более старые версии “Битвы при Равенне” (Lienert 2015). Данная поэма также анонимна. Ранние исследователи предполагали, что и “Бегство Дитриха”, и “Битва при Равенне” были написаны одним автором, однако формальные и стилистические различия между двумя поэмами заставили отказаться от этой теории. Тем не менее, из состояния рукописи становится ясно, что современники рассматривали два эпоса как единое произведение (Millet 2008). Для “Битве при Равенне” основной редакцией считается та же, что и для “Бегства Дитриха” (редакция R), поэтому для корпуса тоже выбрана она.

“Битва при Равенне” состоит из 1140 строф, форма которых не встречается ни в одной другой поэме. Как и другие героические эпосы, эта поэма, вероятно, предназначалась для устного исполнения, но мелодия не сохранилась. (Heinzle 1999) анализирует строфу следующем образом: она состоит из трех “длинных строк” с рифмами на цезурах. Первая строка состоит из трех метрических стоп перед цезурой, затем трех дополнительных стоп; вторая – из трех стоп перед цезурой, затем четырех дополнительных стоп; и третья – из трех стоп перед цезурой, и пять или даже шесть дополнительных стоп после.

В поэме содержатся многочисленные намеки на “Песнь о Нибелунгах”, начиная с вступительной строфы, в которой цитируется вступительная строфа “Песни о Нибелунгах” из редакции C. (Haymes, Samples 1996) предполагают, что “Битва при

Равенне” существует как своего рода приквел к “Песни о Нибелунгах”. (Curschmann 1989) считает, что встречи Дитриха и Зигфрида здесь и в *Rosengarten* восходят к устной традиции. Однако (Lienert 1999) рассматривает сражения в “Битве при Равенне” как часть литературного соперничества между двумя традициями.

2.3.4. Alpharts Tod “Смерть Альфарта”

“Смерть Альфарта” рассказывает о юном герое Альфарте, одном из героев Дитриха и племяннике Гильдебранда, в начале войны между Дитрихом и его дядей Эрменрихом. Альфарт настаивает на том, чтобы отправиться в путь в одиночку, и, хотя он храбрый и могущественный воин, в конце концов он сталкивается с Витеге и Хайме, двумя предателями, перешедшими на сторону Эрменриха. Они убивают его бесчестным образом; Эрменриху тем временем не удается победить Дитриха.

Как и предыдущие поэмы, “Смерть Альфарта” является анонимной. Данный эпос обычно датируется второй половиной XIII века, но его стиль предполагает, что это может быть более новая версия более старой поэмы (Heinzle 1999). Более старая версия скорее всего была написана после “Бегства Дитриха” (Hoffmann 1974). Возможно, текст был написан в Австрии (Lienert 2015). “Смерть Альфарта” представлен в виде единственной бумажной рукописи XV века, которая в XVIII веке была разделена на три части. Сохранилось только ¾ рукописи (Heinzle 1999). В своем незавершенном виде поэма состоит из 469 строф по 4 строки. По строению строфы выглядят так же, как и в “Песни о Нибелунгах”.

“Смерть Альфарта” предлагает своего рода альтернативу событиям, описанным в поэме “Бегство Дитриха”: вместо того, чтобы пригласить Дитриха навестить его с намерением убить, Эрменрих открыто объявляет войну. И вместо того, чтобы добиться успеха в изгнании Дитриха, Эрменрих терпит неудачу. В поэме сохранены темы взаимоотношений между лордом и вассалом, которые можно найти в “Бегстве Дитриха” и “Битве при Равенне”, но при этом Дитрих изображен опытным и успешным. Тем не менее, его враги убегают, что оставляет возможность для дальнейшего развития событий (Lienert 2015).

2.3.5. Dietrich und Wenezlan “Дитрих и Венецлан”

“Дитрих и Венецлан” рассказывает о Венецлане из Польши, который захватил в плен одного из воинов Дитриха, тем самым бросая ему вызов. Дитрих находится при дворе Этцеля, когда Вольфхарт, который вместе с Хильдебрандом был схвачен Венецланом, прибывает, чтобы сказать, что Венецлан хочет сразиться с Дитрихом один на один – если Дитрих победит, то Венецлан освободит Вольфхарта и Хильдебранда. Поначалу Дитрих сопротивлялся, но когда Вольфхарт разозлился и обвинил Дитриха в трусости, сказав, что если Дитрих откажется, Венецлан нападет на Этцеля с армией, Дитрих сказал, что пошутил, и, конечно же, будет сражаться, чтобы освободить своих вассалов. Далее в рукописи следует пробел. А после поэма продолжается битвой между Дитрихом и Венецланом. Фрагмент обрывается без развязки.

Данный эпос сохранился только в единственном, неполном и фрагментированном варианте из примерно 500 рифмованных двустиший. Неясно, был ли этот фрагмент главным в поэме или отдельным эпизодом из более длинной поэмы. Сам текст, возможно, был написан около 1250 года. Предполагается что он был написан на десятилетия раньше, чем другие исторические поэмы из цикла текстов про Дитриха (Heinzle 1999).

Поэма лишь отчасти вписывается в категорию исторических эпосов о Дитрихе, хотя оно явно относится к ситуации изгнания, описанной в “Бегстве Дитриха” и “Битве при Равенне”. (Millet 2008) считает, что данный текст нельзя отнести ни к одной из групп, потому что поэма слишком фрагментирована. Вызов на сражение больше напоминает фантастические поэмы. Первоначальный отказ Дитриха сражаться и обвинение его в трусости также имеют больше общего с фантастическими поэмами, где такие события происходят часто. Lienert (2015), таким образом, описывает “Дитриха и Венецлана” как нечто среднее между двумя группами эпосов о Дитрихе.

Венецлан из Польши (*Wenezlan von Bôlân*), возможно, вдохновлен Венцеславом I, герцогом Богемии. Однако Венецлан – король Польши, а его воины на самом деле русские (*Riuzen*), что ослабляет связь с Венцеславом I. (Millet 2008) соглашается с

(Eis 1953) в том, что источником поэмы может быть нижненемецкий устный рассказ о битвах Дитриха против славян. Данная теория основывается на наличии похожих битв в “Саге о Тидреке” (*Thidrekssaga*), древнескандинавской саге. В некоторых из этих битв проскальзывают параллели с пленением Вольфхарта.

2.3.6. Rosengarten zu Worms “Розовый сад в Вормсе”

В поэме рассказывается о битве между героями цикла текстов о Дитрихе и персонажами “Песни о Нибелунгах”, которая происходит в розовом саду в городе Вормс. Драка вызвана желанием Кримхильды испытать характер своего жениха Зигфрида в поединке с Дитрихом Бернским. В конце концов, Дитрих и его воины побеждают бургундов, включая Зигфрида.

“Розовый сад”, возможно, был написан еще до 1250 года, отчасти из-за своей тесной связи с поэмой *Biterolf und Dietleib*. Однако достоверно поэма датируется примерно 1300 годом. Где и кем могла быть написана поэма, неясно. Обычно выделяют пять основных версий “Розового сада”: A, DP, F, C и средненижненемецкий вариант. Версия A содержит около 390 строф и может быть дополнительно подразделена на более старую, более новую и дрезденскую версии. Вариант DP может быть дополнительно подразделен на версии D, содержащей около 633 строф, и P, причем P является сокращенной версией D. Версия F является фрагментарной, в то время как C представляет собой смесь A и DP (Heinzle 1999). Для нашего корпуса была выбрана версия A. Строфа, используемая в “Розовом саду”, построена также, как и в “Песни о Нибелунгах”, все строки имеют одинаковую длину. Каждая строка состоит из трех метрических стоп, цезуры и трех дополнительных стоп (Millet 2008).

Обычно эту поэму рассматривают как осуждение Кримхильды и, как следствие, ее роли в “Песни о Нибелунгах”. “Розовый сад” часто воспринимают как металитературный текст, в котором обсуждается природа героической поэзии. Данный эпос содержит многочисленные цитаты и намеки на другие героические поэмы с помощью названий таких предметов, как мечи, а также имен различных героев и лошадей. Более того, версия D цитирует или переворачивает многие мотивы исторических эпосов о Дитрихе, а именно события из “Бегства Дитриха” и “Битвы

при Равенне". В "Розовом саду" Этцель ищет Дитриха, тем самым ссылка Дитриха при дворе Этцеля отменена. Кроме того, текст содержит явный намек на роль Витеге в "Смерти Альфарта" (Lienert 2015).

Несмотря на то, что "Розовый сад" сочетает в себе персонажей из "Песни о Нибелунгах" и цикла текстов о Дитрихе, его обычно считают одной из мифических поэм о Дитрихе. Что касается устной традиции, то примечательно, что отец Кримхильды и Гюнтера носит имя Гибих, соответствующее скандинавской традиции, которое в "Песне о Нибелунгах" было заменено другим именем. Это говорит о том, что поэт знал устную традицию, независимую от письменной "Песни о Нибелунгах", хотя автор явно ссылается на различные аспекты этой поэмы и даже цитирует различные строки (Heinzle 1999).

2.3.7. Wunderer "Чудовище"

В "Чудовище" рассказывается о встрече молодого Дитриха с Чудовищем, когда Дитрих гостил при дворе Этцеля. Чудовище охотится и хочет съесть девушку, которая, как позже оказывается, является госпожой Удачей (*Frau Saelde*). Дитрих защищает ее от Чудовища и получает ее благословение после своей победы.

Возможно, этот эпос был написан еще в XIII веке, но впервые он засвидетельствован только в XV веке. Поэма также является анонимной. Разные версии "Чудовища" написаны рифмованными двустишиями и строфами. Для нашего корпуса взята версия *Dresdener Heldenbuch*, состоящая из строф. Строфа состоит из четырех "длинных строк". Каждая строка состоит из трех стоп, цезуры и трех дополнительных стоп, как и в "Песни о Нибелунгах". Слово перед цезурой рифмуется со словом перед цезурой в следующей строке (Heinzle 1999). (Hoffmann 1974) также интерпретирует данную строфу как состоящую из восьми коротких строчек с чередующимися рифмами.

"Чудовище" часто считается необычным произведением среди мифических поэм о Дитрихе. По длине она больше напоминает балладу, чем типичный героический эпос (Millet 2008). Дитрих не находится в изгнании при дворе Этцеля, а скорее был отправлен туда для получения образования, что тоже больше свойственно балладами, и он находится там в более молодом возрасте, чем в остальных эпосах.

Более того, в отличие от большинства мифических поэм о Дитрихе, Дитрих не проявляет неохотности или трусости, а скорее, наоборот, горит желанием помочь нуждающейся девушке (Hoffmann 1974).

2.4. Модели для работы с историческими текстами

В ходе экспериментов с несколькими моделями было выбрано три, которые были дообучены на средневерхненемецкой части датасета (Besnier и Mattingly 2021). Первая модель – “atzenhofer/distilroberta-base-mhg-charter-mlm” на Hugging Face[1], предназначенная для задачи заполнения масок и дообученная на средневерхненемецких хартиях датасета Monasterium.net. Monasterium.net дает доступ ко множеству исторических документов из европейских архивов (Monasterium.net 2024). Он был разработан международной ассоциацией архивов и научно-исследовательских институтов ICARus с целью создания общей инфраструктуры для хранения и изучения хартий (Heinz 2015). Эта платформа содержит более полумиллиона средневековых и ранних современных документов, они являются ценным ресурсом для историков, лингвистов и многих других исследователей. Ресурс содержит множество хартий на средневерхненемецком языке.

RoBERTa имеет ту же архитектуру, что и BERT, но токенизирует текст на байтовом уровне (аналогично GPT-2) и использует другую схему предобучения. Оригинальные модели BERT и RoBERTa достаточно велики и ресурсоемки, что затрудняет их обучение, поэтому авторы данной модели решили использовать более легкую версию. DistilRoBERTa, облегченная версия RoBERTa, обучена предсказывать те же замаскированные слова, что и оригинальная RoBERTa. Облегчена она за счет “дистилляции знаний” (Knowledge Distillation, KD). Это подход к сжатию модели, при котором модель меньшего размера обучается имитировать более крупную, предварительно обученную модель. В статье (Sahn et al. 2019), которая посвящена DistilBERT, авторы также удалили из модели некоторые функциональные возможности, а именно встраивание типа токена, которое определяет принадлежит ли токен к первому или второму предложению, и pooler, который создает единое

[1] https://huggingface.co/atzenhofer/distilroberta-base-mhg-charter-mlm

векторное представление для всего входного сигнала, чтобы еще больше снизить вычислительную нагрузку. Другим важным моментом является введение “тройной функции потери” во время обучения, которая включает в себя перекрестную энтропию (cross-entropy), маскированное языковое моделирование (masked language modeling) и косинусное расстояние между эмбеддингами (cosine embedding loss). Авторы утверждают, что каждый компонент “тройной функции потери” необходим для достижения наилучших результатов. Облегченная версия продемонстрировала производительность на 97% выше, чем у более крупной модели, но при этом она была на 40% меньше и на 60% быстрее. DistilRoBERTa получается из оригинальной RoBERTa по аналогичной схеме.

Вторая используемая в нашей работе модель, “christinbeck/GHisBERT” на Hugging Face[2], также изначально обучена для задачи заполнения масок. Основная цель работы (Beck, Köllner 2023) – это создание модели, с помощью которой можно было бы эффективно проводить лексико-семантические исследования на разных исторических этапах немецкого языка. За основу взята модель BERT, которая была дообучена. Для обучения использовались данные из двух источников: собрание корпусов Reference Corpora of Historical German (Zeige et al. 2022), в котором находятся корпуса для древневерхненемецкого, средневерхненемецкого, ранненововерхненемецкого и средненижненемецкого языков, и, чтобы сбалансировать обучающие данные, авторы взяли данные на современном немецком языке из Deutsches Textarchiv (Deutsches Textarchiv 2023). В результате авторы статьи показывают, что обученная с нуля на соответствующих исторических данных модель BERT обеспечивает более высокие результаты, чем просто дообученная модель на какое-либо необходимое задание.

Третья используемая модель – это модель на базе Python-библиотеки Flair[3], разработанная исследователями из Берлинского университета имени Гумбольдта. Ее основная идея заключается в предоставлении простого и унифицированного интерфейса для различных NLP задач (Akbik et al. 2019). Данный интерфейс также

[2] https://huggingface.co/christinbeck/GHisBERT
[3] https://github.com/flairNLP/flair/

предоставляет стандартные процедуры обучения модели и выбора гиперпараметров, а также легкого использования дополнительных данных, которые можно загрузить из публично доступных NLP датасетов и преобразовать их в необходимый для обучения модели вид. Кроме того, Flair уже предоставляет небольшую коллекцию предобученных моделей и не только. Данная библиотека включает в себя широкий спектр эмбеддингов, включая эмбеддинги контекстуализированных и обычных слов, пар байтов и символов. Авторы разработали простые базовые классы, которые предоставляют общий функционал NLP, и создали отдельные классы для классических эмбеддингов, таких как GloVe и fastText, а также для эмбеддингов ELMo, BERT, LSTM и предложенных создателями эмбеддингов Flair. Еще одним из удобств является возможность использовать предобученные модели, основанные на архитектуре трансформеров, например, тот же BERT, GPT-2 и многие другие, то есть эмбеддинги не привязаны к конкретной модели, как это обычно бывает. Для задачи классификации последовательности или текстов также предоставляются уже готовые архитектуры, например, двунаправленная LSTM и условные случайные поля (Conditional Random Fields – CRF), а также можно обучить собственные эмбеддинги и модели, необходимые для специфических индивидуальных задач. Flair используют для таких прикладных задач, как распознавание именованных сущностей, разметка частей речи и различных видов классификации текста.

В Flair уже есть предобученная NER модель для немецкого языка, но она использует не подходящие для нашей задачи теги, а именно PERSON, LOCATION, ORGANISATION и MISCELLANEOUS. Интересно, что в этой модели используется эмбеддинги XLM-RoBERTa и новый подход FLERT, который позволяет модели учитывать контекст на уровне документа, хотя обычные NER модели в основном работают на уровне предложений. Для этого в каждое предложение, которое необходимо классифицировать, включается по 64 токена правого и левого контекстов. Благодаря включению информации на уровне документов, устраняются несоответствия в классификации объектов в документе и повышается производительность модели. Модели с FLERT показали лучшие результаты по

сравнению с дообученными трансформерами на двух языках – английском и немецком (Schweter, Akbik 2021).

Для нашей работы была использована модель Flair, направленная на разметку последовательностей, и для нее были инициализированы эмбеддинги Flair для современного немецкого языка. В статье (Akbik et al. 2018) представлен новый тип эмбеддингов для платформы Flair. Авторы называют их “контекстные эмбеддинги строк” (contextual string embeddings), в которых каждая последовательность символов может быть представлена по-разному, в зависимости от контекста на уровне предложения. Такие эмбеддинги генерируются с помощью BiLSTM: языковая модель учится предсказывать следующий символ в последовательности символов, побочным эффектом чего является то, что BiLSTM запоминает контекстуализированные представления последовательности символов. Таким образом, данные эмбеддинги могут охватить скрытую синтаксическую и семантическую информацию, что особенно полезно для задачи классификации последовательности. Эксперименты авторов с предлагаемыми эмбеддингами показывают значительные улучшения во многих задачах NLP, например, в статье описываются эксперименты по распознаванию именованных сущностей для английского и немецкого языков, а также на разметке частей речи и разбиении фраз. Для удобства авторы интегрировали данные эмбеддинги в библиотеку Flair. В нашем исследовании данная модель будет единственной из представленных, которая изначально обучена на современном немецком языке.

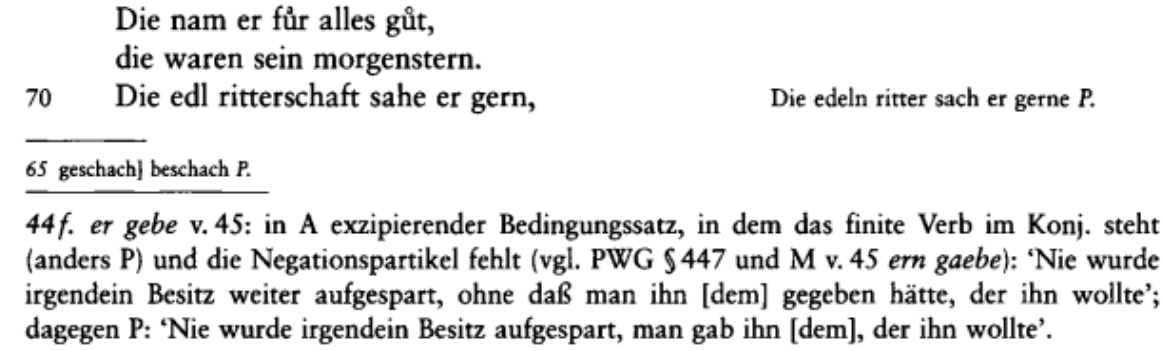
3. Практическая часть

3.1. Сборка и обработка корпуса

Семь выбранных для корпуса текстов изначально были представлены в PDF формате: *Alpharts Tod* (Lienter, Meyer 2007), *Dietrich und Wenezlan* (Lienter, Meyer 2007), *Dietrichs Flucht* (Lienert, Beck 2003), *Nibelungenlied* (Reichert 2005), *Rabenschlacht* (Lienert, Wolter 2005), *Rosengarten* (Lienert et al. 2015), *Wunderer* (Kragl 2015). Такой формат является не самым удобном для считывания текста компьютерами. Для их перевода в текстовый формат использовались

Python-библиотеки PDFMiner (Valvekens 2023) и PyPDF2 (Fenniak 2024). PDFMiner использовалась для шести текстов и только *Wunderer* был обработан библиотекой PyPDF2, так как кодировка файла не распознавалась правильно другой библиотекой. Но даже специальные библиотеки не всегда все правильно опознавали или не всегда считывали только то, что нужно. Во всех файлах присутствовали сноски, в которые выносилась дополнительная информация о тексте. Отрывок исходного текста “Бегство Дитриха” с примером умлаутов и сносок можно увидеть на рисунке 1. Чтобы очистить основной текст от них и поправить некоторые особые знаки, применялись регулярные выражения.

Рисунок 1. Пример исходного фрагмента текста “Бегство Дитриха” (Lenter, Wolter 2003)


Die nam er fůr alles gůt,
die waren sein morgenstern.
70 Die edl ritterschaft sahe er gern, Die edeln ritter sach er gerne *P.*

65 geschach] beschach *P.*

44f. er gebe v. 45: in A exzipierender Bedingungssatz, in dem das finite Verb im Konj. steht (anders P) und die Negationspartikel fehlt (vgl. PWG § 447 und M v. 45 *ern gaebe*): ‘Nie wurde irgendein Besitz weiter aufgespart, ohne daß man ihn [dem] gegeben hätte, der ihn wollte’; dagegen P: ‘Nie wurde irgendein Besitz aufgespart, man gab ihn [dem], der ihn wollte’.

В первую очередь из всех текстов удалялись цифры, будь то нумерация строк, начало сноски к конкретной строке или нумерация страниц. Перед нумерацией страниц также часто присутствовала буква, которую необходимо было захватить. Еще одним общим элементом очистки являлись остатки кодировки, например, “\x0c” или “\xa0”. Это встречалось в большинстве текстов. Кроме того, даже с помощью специальных библиотек плохо распознавались умлауты, особенно в более старых вариантах текстов. Также из каждого текста удалялись верхние колонтитулы с названием текста. Обе версии “Чудовища” были представлены в одном файле через страницу, то есть на первой странице первая версия (Dresdener Heldenbuch):

(1) *[242v] Konick Etzel sprach zu hande: ›so heiß sie komen her!‹*
der pfortner sie pald fande vnd saget ir die mer. (Kragl 2015)

на второй – вторая (Straßburger Druck):

(2) *Künig Etzel sprach behende: ›so heiß sie kommen her!‹*
der thorwart sie bald fande; er saget ir die mer (Kragl 2015)

На третьей будет продолжение первой версии и так далее.

В результате тексты были по большей части очищены. Самым чистым текстом получилась “Песнь о Нибелунгах”, в том числе версия B, которая была выбрана для корпуса. В файле с текстом не было сносок, все комментарии были указаны после. Кроме того, в нем не было верхних колонтитулов, так что нужно было убрать только нумерацию страниц и строк с буквой версии. В некоторых частях остальных текстов сноски до конца не очистились, потому что невозможно было учесть все варианты в регулярных выражениях, но не очистилась только небольшая часть текстов, поэтому на обучение модели это не повлияло.

Код, использовавшийся для предобработки текстов, опубликован на GitHub[4]. Он разделен на части, каждая часть посвящена конкретному тексту, очищались разные версии одних и тех же текстов, то есть для корпуса уже потом были отобраны только 7 текстов из 22 очищенных.

Статистику по количеству предложений и словоупотреблений в корпусе можно увидеть в таблице 1. В ней представлено число предложений, словоупотреблений и процент словоупотреблений по отношению ко всему корпусу для каждого текста. Заметно, что “Песнь о Нибелунгах” включает в себя больше всего словоупотреблений, но процент словоупотреблений составляет чуть больше 30%, поэтому по большей части корпус состоит из текстов, которые ранее не включались в другие датасеты.

[4] https://github.com/Maryleya/NER-in-medieval-German-texts/blob/main/PDFtoText.ipynb

Таблица 1. Статистика по собранному корпусу.

Тексты	Число предложений	Число словоупотреблений	Процент словоупотреблений
Alpharts Tod	913	16 813	6,5%
Dietrich und Wenezlan	179	2 707	1,1%
Dietrichs Flucht	4 215	68 125	26,5%
Das Nibelungenlied	2 652	83 817	32,6%
Die Rabenschlacht	2 792	43 008	16,7%
Rosengarten zu Worms	1 982	34 332	13,4%
Der Wunderer	210	7 965	3,1%
Всего	**12 943**	**256 767**	**100%**

3.2. Составление датасета

Для проверки качества обученных моделей был вручную составлен датасет именованных сущностей из текстов о Дитрихе на основе каталогов наименований (Namensverzeichnis) из пяти текстов – “Смерть Альфарта” (Lenter 2007), “Дитрих и Венецлан” (Vollmer-Eicken 2007), “Бегство Дитриха” (Lenter, Wolter 2003), “Битва при Равенне” (Lenter 2005) и “Розовый сад” (Lienert et al. 2015). Единственный текст из произведений о Дитрихе, для которого такого каталога нет, это “Чудовище”. Но он является достаточно маленьким по объему, и практически все именованные сущности, встречающиеся в этом тексте, уже входят в датасет, потому что они появляются и в других текстах про Дитриха, поэтому “Чудовище” также учитывалось при дальнейшей оценке качества моделей. Кроме того, в наш датасет были включены именованные сущности для текста “Песнь о Нибелунгах” из датасета (Besnier, Mattingly 2021).

В каталог наименований входят все, встречающиеся в данном тексте, имена людей и названия народов, наименования географических мест, а также имена различных вещей, принадлежащих героям, например, мечей и лошадей. Каждое имя

сопровождается объяснением, кто это, а каждое место – описанием, где оно находится, это может быть определено точно или предположительно. Кроме того, указывается номер строки, в которой упоминается данная сущность. Также в каталоге даны варианты наименований сущностей в других текстах о Дитрихе, иногда предположительные, часто варианты отличаются друг от друга только чередованием одной или двух букв, но иногда сущности могут быть совсем не похожи друг на друга, но обозначать одно и то же в разных текстах.

На основе этой информации был собран датасет, который состоит из трех колонок: lemma, tag, tokens. В колонке lemma указывалась сущность в начальной форме, если возникала какая-то вариативность между каталогами, то бралась та форма, которая упоминалась в большем количестве текстов, но таких спорных случаев практически не возникало.

Что касается колонки tag, то в ней указывался один из следующих тегов: PERSON, REGION, CITY, WATER, GROUP или PROPERTY. Тег PERSON был присвоен именованным сущностям, которые обозначают отдельных людей в текстах. Он относится к любому персонажу или историческому лицу, упомянутому в этих средневерхненемецких текстах, независимо от их роли и значимости в повествовании. Пример – Dietrich, Kriemhild, Seyfrid. Тег PLACE, который был использован в датасете (Besnier, Mattingly 2021), в нашем датасете было решено разделить на три подтега, а именно REGION, CITY и WATER. Тег WATER использовался для именованных объектов, которые обозначали моря, реки, озера или любой другой водоем. Пример – Donau, Main, Reine. Тег CITY присваивался именованным сущностям, обозначающим городские поселения с четко очерченными границами, которые обычно обозначают центры проживания людей. Пример – Bern, Ilsenburg, Kiew. Тег REGION использовался для обозначения более крупных географических районов и территорий, таких как провинции, страны или континенты. Пример – Bayern, Dänemark, Engelant. Тег GROUP был присвоен именованным сущностям, которые в совокупности обозначали несколько человек. Обычно это названия народов, жители конкретных городов или регионов, но также он присваивался любым другим объединениям людей, упомянутым в текстах.

Пример – Dänen, Franncken, Nibelungen. И, наконец, тег PROPERTY использовался для обозначения именованных сущностей, которые в тексте относились к материальным благам или любому имуществу, которым владеют личности. Это могли быть наименования конкретных предметов, особенно относящихся к броне, например, меч или шлем, или что-либо другое, представляющее материальную ценность, а также в этот тег попадали имена животных, принадлежащих персонажам, а конкретнее имена лошадей. Пример – Falcke, Gleste, Mimming.

В таблице 2 представлена статистика по каждому тегу. Общее количество лемм в датасете составило 387. Самым большим по количеству лемм тегом получился PERSON. Самым же маленьким получился тег WATER, а следом за ним идут GROUP и PROPERTY. В текстах данные категории упоминаются редко. Для работы с моделями некоторые теги были объединены, иначе бы модель просто не смогла бы обучиться на таком маленьком количестве данных.

Таблица 2. Статистика по тегам

Тег	Количество лемм
PERSON	223
REGION	75
CITY	64
WATER	7
GROUP	9
PROPERTY	9
Всего	**387**

Колонка tokens включает в себя все варианты определенной леммы. Во-первых, это чередования конкретных букв. В частности, во всех каталогах указываются следующие пары:

- B – P (*Balmung – Palmung*)
- C – K (*Criemhild – Kriemhild*) и иногда C – CH (*Zazamanc – Zazamanch*)

- I – Y (*Crist – Cryst*) и иногда I – E (*Ilsam – Elsan*)
- J – G (*Jorg – Georg*)
- F – V (*Falcke – Valche*)

Во-вторых, к токенам для некоторых лемм относятся их словоформы в разных падежах и числах, так как в текстах могли встречаться разные варианты, например, *Mimming – Mimmingen – Mimminges*. В-третьих, в токены входят разные варианты написания или распознавания умлаутов и других фонетических явлений:

- с умлаутом – без умлаута (*Österreich – Osterreich*)
- с умлаутом – другое написание умлаута (*Götelind – Goetelind*)
- разные варианты написания дифтонгов и умлаутов (*Zæringen – Zeringen*)

Также встречаются и некоторые другие варианты токенов, например, удвоение или пропуск буквы. В данный датасет мы постарались включить все возможные варианты токенов, встречающиеся в текстах, но стоит учитывать тот факт, что некоторые токены при переводе в текстовый формат могли распознаться некорректно, поэтому при оценке моделей на сущностях из данного датасета стоит брать во внимание небольшую погрешность. Все токены для каждой леммы отсортированы по алфавиту и написаны через знак-разделитель – ":".

Таким образом, в результате работы имеется датасет из 387 лемм именованных сущностей из семи текстов собранного корпуса, которые разделены на 6 разных тегов. Ниже в таблице 3 представлен пример трех полных строчек из датасета.

Таблица 3. Пример собранного датасета

lemma	**tag**	**tokens**
Azagouc	REGION	Azagouch
Badowe	CITY	Badaw:Badawe:Badua:Batowe
Baldewein	PERSON	Baldewin:Baldwin

3.3. Методология

3.3.1. Подготовка данных

Модели будут обучаться на именованных сущностях из текста “Песнь о Нибелунгах”, которые взяты из датасета (Besnier, Mattingly 2021). Для обучения мы взяли три тега, из которых изначально состоял датасет, – PERSON, PLACE и GROUP, потому что из таблицы 3 видно, что при разделении тегов количество сущностей в них становится совсем маленьким, и модели не смогут хорошо обучиться на таких данных.

Для начала необходимо провести BIO-разметку. BIO-аннотации являются популярным методом разметки данных в моделях машинного обучения, особенно для задач распознавания именованных сущностей. BIO означает Beginning (начало), Inside (внутри), Outside (вне). Эти теги используются для обозначения грани сущностей в тексте и назначаются следующим образом. Тег B обозначает начало сущности. Если сущность состоит из одного слова, то этот тег присваивается ему, а если из более чем одного слова, то этот тег используется только для первого слова. Тег I используется для всех слов, найденных внутри сущности после первого слова. Если сущность содержит только одно слово, то тег I не будет использоваться для нее. Тег O используется для всех слов, которые не являются частью сущности. Для тегов B и I через знак “-” указывается собственно тег, к которому данная сущность относится. Например, рассмотрим предложение “*Dö kom der herre Sivrit da er die frouwen sach.*” Теги в нем будут выглядеть так:

- Dö (O)
- kom (O)
- der (O)
- herre (O)
- Sivrit (B-PERSON)
- da (O)
- er (O)
- die (O)
- frouwen (O)

- sach (O)
- . (O)

Таким образом, BIO-разметка обеспечивает структурно согласованный способ обучения моделей машинного обучения таким задачам, как распознавание именованных сущностей, позволяя им понимать и предсказывать не только тип сущности, но и его размер в пределах предложения. Это упрощает процесс обучения, а также приводит к повышению производительности моделей.

При ограниченном наборе данных становится крайне важным максимально использовать каждое значение. Так, если предложение полностью состоит из тегов O без каких-либо сущностей, это не вносит существенного вклада в обучение модели распознавания сущностей. Основная цель обучения NER модели – правильно идентифицировать и классифицировать именованные сущности в заданных данных. Предложения, состоящие исключительно из тегов O имеют ограниченную ценность для достижения этой цели. Они могут непреднамеренно привести к смещению модели в сторону предсказания тега O. Следовательно, в условиях, когда данных мало, может быть полезно исключить такие предложения, чтобы быть уверенным в том, что модель учиться только на полезных обучающих примерах, которые включают в себя сущности.

Датасет (Besnier, Mattingly 2021) является относительно маленьким, поэтому при BIO-разметке текста из него были убраны все предложения, которые состоят только из тега O. Цель состоит в том, чтобы включить достаточное разнообразие тегов для эффективного обучения моделей, но при этом не снижать качество. В предложениях с сущностями действительно хватает примеров на тег O, а если добавлять еще предложения без тегов, то у модели качество падает.

3.3.2. Обучение моделей

Параметры обучения модели Distilroberta-base-mhg-charter-mlm:

- количество эпох – 20
- оптимизатор – AdamW
- learning rate (скорость обучения) – 1e-4

- weight decay (сокращение веса) – 0,01

Для токенизации данных использовался RobertaTokenizer.

Параметры обучения модели GHisBERT:

- количество эпох – 20
- оптимизатор – AdamW
- скорость обучения – 1e-4
- сокращение веса – 0,01

Для токенизации данных использовался BertTokenizerFast.

Параметры обучения модели Flair:

- количество эпох – 20
- оптимизатор – SGDW
- скорость обучения – 1e-1

Для токенизации данных использовался segtok.

3.3.3. Оценка моделей

В машинном обучении важно, чтобы модель не просто запоминала обучающие данные, а изучала закономерности и шаблоны, которые после могут быть замечены в данных, которые модель не видела. Это связано с тем, что в реальности данные, с которыми мы сталкиваемся, часто отличаются от данных, на основе которых была обучена модель. Поэтому важно знать, насколько хорошо наша модель может правильно присваивать теги новым данным, которые она ранее не видела. Более того, это может помочь выявить любые проблемы с переобучением. Когда модель переобучена, она очень хорошо работает на обучающих данных, но плохо на незнакомых, потому что она слишком тщательно изучила обучающие данные, включая шумы и выбросы, без учета общих закономерностей.

С помощью собранного нами датасета мы разметили тексты о Дитрихе. Теги WATER, REGION и CITY были возвращены к единому тегу PLACE, а тег PROPERTY был включен в тег PERSON, потому что дообученные модели умеют распознавать только три тега – PERSON, PLACE и GROUP. Затем мы применили к текстам о Дитрихе все три модели, которые обучались только на тексте “Песнь о Нибелунгах”.

Эпосы про Дитриха модели не видели, поэтому мы получим более реалистичную оценку их производительности.

Для оценки использовалась библиотека seqeval (Nakayama 2018), которая часто используется для задач классификации последовательностей. Благодаря clssification report (отчет о классификации) в seqeval возможно получить гораздо более четкое и детальное представление о том, как работает модель для каждого класса в отдельности, что как раз важно для задачи мультиклассовой классификации. Отчет о классификации поддерживает BIO-разметку, что отлично подходит для распознавания именованных сущностей, так как он сможет оценивать каждую размеченную сущность, а не отдельные токены. Чтобы сгенерировать отчет о классификации, необходимо предоставить два списка: один с истинными метками, а другой с предсказанными метками. Оба списка должны иметь одинаковую длину, и каждый список содержит список тегов для каждого предложения. Отчет о классификации предоставляет информацию о precision (точность), recall (полнота) и F1-score (F1-мера) для каждого класса в отдельности, в нашем случае для PERSON, PLACE и GROUP. Кроме того, он выводит micro-average (среднее значение на микроуровне), macro-average (среднее значение на макроуровне) и weighted-average (среднее взвешенное значение) для точности, полноты и F1-меры. Среднее значение на микроуровне рассчитывает результат с учетом общего количества true positives (истинно положительные), false negatives (ложноотрицательные) и false positives (ложноположительные). Среднее значение на макроуровне рассчитывает результат для каждого тега и находит их невзвешенное среднее значение, то есть без учета количества каждого тега в датасете. Среднее взвешенное значение рассчитывает метрики для каждого тега и выводит среднее значение с учетом количества каждого тега.

Последний показатель часто используется для оценки эффективности моделей, направленных на мультиклассовую классификацию, особенно в задачах, где присутствует дисбаланс данных. В таких задачах на одни классы будет намного больше примеров, чем на другие. Если использовать среднее значение на микро- или макроуровне, то эффективность классификатора в классе с наибольшим количество

данных будет преобладать над средним показателем, что потенциально скроет низкую эффективность в классе с наименьшим количеством данных. Среднее взвешенное значение придает большее значение классам с более высоким количеством данных и меньшее – тем, у которых меньше данных. Этот показатель дает более точное представление о результатах работы модели среди всех классов, гарантируя, что результаты работы модели в классе с маленьким количеством данных пропорционально повлияют на общий балл, так что результаты, при необходимости, отразят низкую эффективность . Поэтому в нашей работе в результатах приводится именно среднее взвешенное значение.

Что касается метрики accuracy (аккуратность), то его использование для мультиклассовой классификации часто приводит к ошибочным выводам. В частности при несбалансированном распределении по классам модель может достичь высокой accuracy за счет того, что наиболее часто встречающиеся классы предсказываются правильно, в то время как менее распространенные классы часто ошибочны. Так, например, происходит с этими моделями, метрика accuracy достигает таких высоких показателей за счет того, что тег O часто находится правильно, поэтому не стоит полагаться на аккуратность, лучше опираться на F1-меру, которая учитывает точность и полноту.

3.4. Сравнение результатов работы моделей и анализ ошибок

3.4.1. Distilroberta-base-mhg-charter-mlm

Дообученная модель показала результаты, представленные в таблице 4. Построенная матрица ошибок для нее представлена на рисунке 2. Метрика accuracy равна 0,94.

Distilroberta-base-mhg-charter-mlm неплохо распознает сущности с тегом PERSON, но все же испытывает затруднения при отличии его от тега O. Например, тегом PERSON модель отмечала некоторые предлоги: *vor*, *mit*. А такие имена, как *Dytwart*, *Cryst* и *Ottnit*, были размечены тегом O.

Тег PLACE распознается достаточно плохо из-за того, что она часто приписывает таким сущностям тег PERSON или еще чаще тег O. Например, такие места, как *Bern* и *Franchrich,* модель определила как PERSON, в то время как имена

Hen и *Wylandes* модель отметила тегом PLACE. Скорее всего второе имя попало в PLACE из-за части “land”, хотя местам с такой частью модель часто приписывала тег O (например, *Niderland*). Кроме того, в этот тег также попадали предлоги и артикли: *der*, *von*.

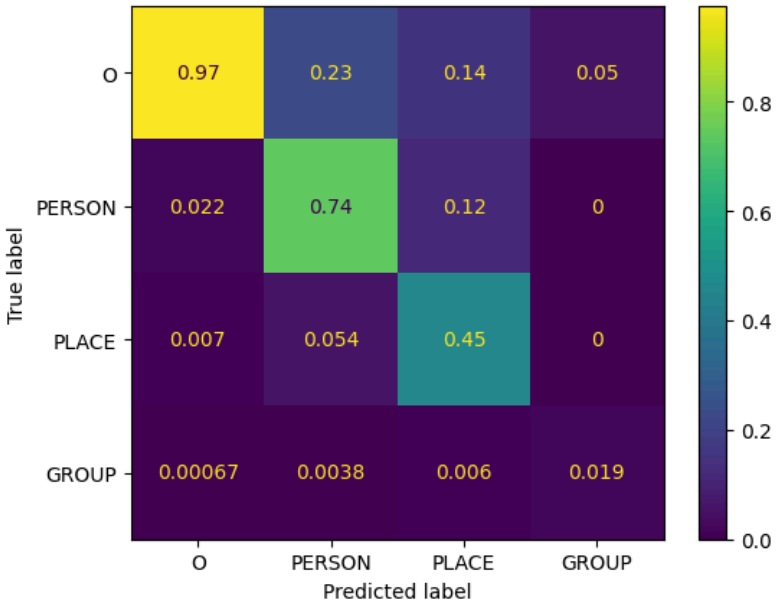
Distilroberta-base-mhg-charter-mlm – единственная модель, которая смогла правильно присвоить тег GROUP хотя бы трем сущностям, хотя остальные сущности с таким тегом отнесла к тегу O (например, *Heunen*). А также она приписала таким группам, как *Amelung* и *Riesen*, теги PERSON и PLACE соответственно.

Эта модель, судя по результатам, занимает среднее положение среди всех трех.

Таблица 4. Результаты модели distilroberta-base-mhg-charter-mlm

	Точность	**Полнота**	**F1-мера**
GROUP	0,27	0,02	0,04
PERSON	0,71	0,74	0,73
PLACE	0,63	0,45	0,53
Среднее взвешенное	0,69	0,68	0,68

Рисунок 2. Матрица ошибок для модели distilroberta-base-mhg-charter-mlm



3.4.2. GHisBERT

Дообученная модель показала результаты, представленные в таблице 5. Построенная матрица ошибок для нее представлена на рисунке 3. Метрика accuracy равна 0,92.

GHisBERT делает еще больше ошибок с тегом PERSON, чем предыдущая, она часто приписывает сущностям с этим тегом тег O. Например, именам *Wolffhart*, *Crimhilt*, *Berhtunch*. А в сам тег PERSON также иногда попадают артикли (*dem*) и глагол *sind*.

Тег PLACE данная модель распознает примерно так же плохо, как и предыдущая, только уже в основном приписывает таким сущностям тег O, а не PERSON. Как и предыдущая модель, GHisBERT разметила несколько сущностей с частью “land”/”lant” тегом O, например, *Grunelant.* А под сам тег PLACE попадали артикли и наречия, например, *der*, *wie*.

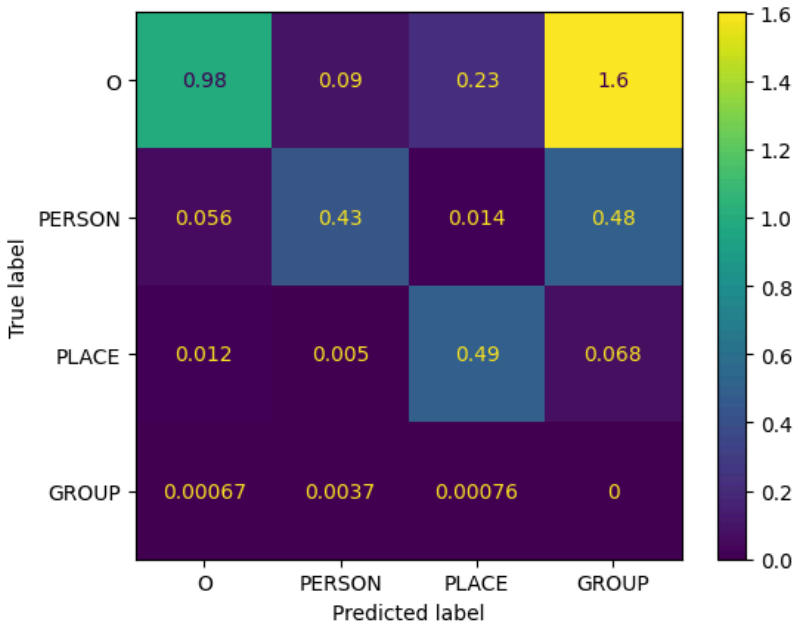
Примечательно, что эта модель очень часто пытается присвоить тег GROUP, но ни разу не получается это сделать успешно, в основном она путает его с тегом O. Под тег GROUP также попадали артикли, предлоги и наречия (*der*, *von*, *also*), а действительные названия групп (*Riuzen*, *Burgonis*, *Hinnen*) были размечены как O. Под тег PERSON, как и в предыдущей модели, попала группа *Amelunch*, а в PLACE оказалось *arabischem*.

Такие показатели говорят о том, что GHisBERT демонстрирует наименее эффективную производительность среди всех моделей.

Таблица 5. Результаты модели GHisBERT

	Точность	**Полнота**	**F1-мера**
GROUP	0,0	0,0	0,0
PERSON	0,81	0,43	0,56
PLACE	0,67	0,49	0,56
Среднее взвешенное	0,79	0,44	0,56

Рисунок 3. Матрица ошибок для модели GHisBERT



3.4.3. Flair

Дообученная модель показала результаты, представленные в таблице 6. Построенная матрица ошибок для нее представлена на рисунке 4. Метрика accuracy равна 0,97.

Flair лучше всего распознает тег PERSON, по сравнению с предыдущими, несмотря на то, что все еще иногда приписывает сущностям с этим тегом тег O. Хотя и встречались некоторые ошибки, например, *Berchtung* и *Saben* были размечены тегом O, а *Heime* и *Wate* тегом PLACE. Но и служебные части речи (*so*, *als*) также попадали в тег PERSON.

Данная модель намного лучше других предсказывает тег PLACE, иногда приписывая таким сущностям тег O и намного реже тег PERSON. В тег O попали такие места, как *Islant* и *Ysterrich*, а в PERSON – *Kerlingen* и *Beren*. Flair намного реже, чем другие две модели, приписывала тег PLACE словам, у которых должен быть тег O (0,04 < 0,14 < 0,23).

В отличие от предыдущих моделей, Flair вообще не предсказывает тег GROUP, и большинство сущностей с тегом GROUP были ошибочно распознаны как O. Под

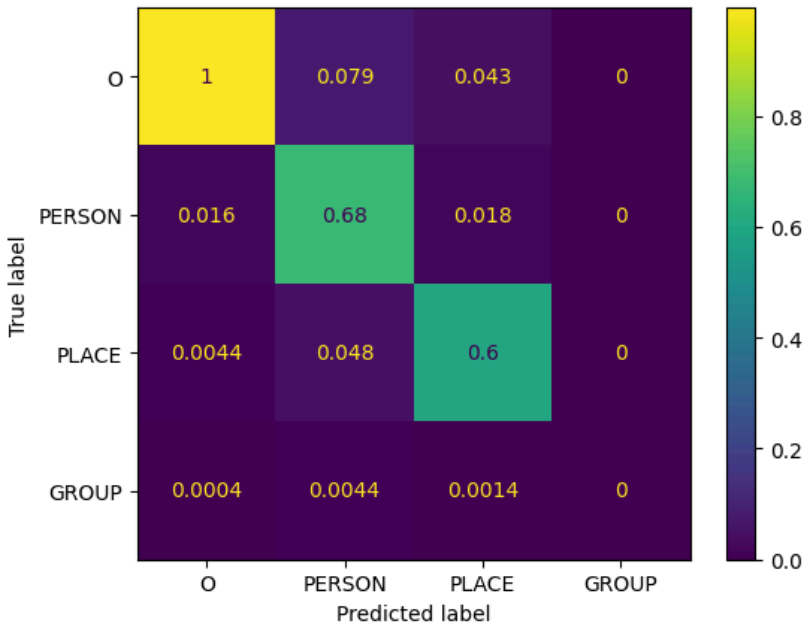
тег O попадали *Amelunch, Hinnen, Riesen*, под тег PERSON – *Amelunge*, под тег PLACE – *Riuzen*.

Тем не менее, несмотря на трудности с тегом GROUP, Flair лучше других моделей распознает тег PERSON и намного лучше распознает тег PLACE, что делает ее наиболее эффективной моделью в данной задаче.

Таблица 6. Результаты модели Flair

	Точность	**Полнота**	**F1-мера**
GROUP	0,0	0,0	0,0
PERSON	0,84	0,68	0,75
PLACE	0,91	0,6	0,72
Среднее взвешенное	0,85	0,65	0,74

Рисунок 4. Матрица ошибок для модели Flair



3.5. Перспективы исследования

Основное направление дальнейшего развития исследования, представленного в данной работе, заключается в улучшении качества работы моделей. В первую

очередь, это можно сделать за счет увеличения количества данных для обучения, что предполагает аннотирование большего количества текстов на средневерхненемецком языке. Как видно из результатов, моделям, как минимум, не хватает данных для тегов PLACE и GROUP, поэтому на обучение необходимо подавать данные с увеличенным количеством именованных сущностей.

Увеличение созданного датасета также является одной из целью дальнейшего развития. Исходя из статистики по тегам, представленной в таблице 2, следует увеличить такие малопредставленные теги, как WATER, GROUP и PROPERTY. Создание более сбалансированного датасета позволит использовать больший объем данных для обучения моделей. Кроме того, улучшить созданный датасет можно за счет введения более разнообразных тегов, по которым потом также можно обучить модели.

Результаты исследования показали, что модель, изначально обученная на современном немецком языке, показывает результаты лучше, поэтому также имеет смысл обратить внимание на другие модели, обученные на современном языке, и сравнить их работу.

4. Практическое применение

Одно из самых популярных направлений применения таких датасетов и, соответственно, моделей, обученных собирать такие датасеты, – это составление графов персонажей. (Kenna, MacCarron 2017) исследуют применение математических графов к древним мифологическим текстам, в частности к эпосам, для анализа персонажей и их взаимодействий в разных произведениях. В начале статьи авторы отмечают, что мифологические эпосы представляют собой сложные системы, состоящие из множества персонажей, событий и отношений между ними. Эти отношения могут быть проанализированы с помощью графов персонажей, которые позволяют выявить скрытые закономерности и взаимосвязи между элементами. Важная часть их работы включает в себя сравнение текстов, в которых происходят исключительно человеческие взаимодействия, и текстов, в которых присутствуют сверхъестественные сущности. С помощью графов можно количественно определить

уровень взаимодействия между персонажами, занимающими схожее положение в текстах. В ходе исследования сравнивались четыре мифологических эпоса. Для них авторы строили графы, где персонажи являлись узлами, а отношения между ними – ребрами. Затем они анализировали эти графы с точки зрения их топологии, плотности, центральности и других характеристик. Результаты анализа показывают, что в мифологических эпосах часто встречаются повторяющиеся паттерны взаимодействия между персонажами. Исследователи считают, что графовый анализ эпосов обеспечивает более глубокое понимание структуры и смысла эпоса. В другой статье (Thuillard et al. 2018) авторы применяют к графам персонажей кластеризацию, классификацию, анализ временных рядов и другие методы компьютерного анализа для исследования структуры мифов на примере нескольких известных мифологических систем.

В статье (Besnier 2020) строятся графы персонажей для тех же трех текстов, что входят и в датасет – *Decem Libri Historiarum*, *Völsunga saga*, *Nibelungenlied*. В средневерхненемецком тексте авторы выделили 67 персонажей, взаимодействия которых они анализировали путем выделения их в отрывке текста. Рассматривался отрывок длиной в три строфы, где персонажи появлялись вместе. Персонажи, которые совсем не взаимодействовали с другими героями не были включены в анализ. Граф “Песни о Нибелунгах” состоял из 50 узлов (персонажей) и 202 ребер (взаимодействий). Затем исследователи вычисляли степень центральности (нормализованное количество персонажей, взаимодействующих с данным персонажем), собственный вектор центральности (показывает влияние персонажа), степень близости (среднее значение кратчайших расстояний между персонажем и всеми другими доступными персонажами) и степень посредничества (количество кратчайших путей, проходящих через определенного персонажа). В результатах показано, что центральная роль Зигфрида возросла в германской традиции по сравнению с другими текстами, так же как и роль Этцеля (Аттилы). Данный анализ показывает сходства и различия между текстами, выявляя, какие персонажи приобрели большее значение, а какие совсем исчезли из повествования, что дает

более полное понимание самих эпосов и того, как повествование в этих текстах изменялось со временем.

Кроме того, датасеты такого типа могут помочь в обогащении аннотаций корпусов общего назначения, например, Middle High German Conceptual Database (MHDBDB) (Zeppezauer-Wachauer 2024), поддерживающаяся Зальцбургским университетом. Данный корпус состоит из средневековой и ранней современной литературы, он предоставляет доступ к наиболее важным произведениям средневерхненемецкой поэзии и множеству других текстов. Тексты хранятся в формате XML-TEI с соответствующими метаданными. А такие корпуса уже открывают путь для более широкого круга исследований средневековых языков и текстов.

5. Заключение

Целью настоящего исследования было собрать и разметить датасет именованных сущностей для текстов на средневерхненемецком языке, а также сравнить несколько NER моделей, обученных на части этого датасета.

В ходе работы были рассмотрены существующие датасеты и корпусы для задачи распознавания именованных сущностей в средневековых языках, исследованы средневерхненемецкие эпосы о Дитрихе Бернском и готовые NER модели для работы с средневековыми языками. Основываясь на полученных сведениях, мы собрали и перевели в текстовый формат корпус, состоящий из семи текстов о Дитрихе Бернском и эпосе “Песнь о Нибелунгах”. На основе этого корпуса мы составили и опубликовали датасет[5] именованных сущностей для обучения на средневерхненемецком языке. В него входит 387 лемм с тегами PERSON, REGION, CITY, WATER, GROUP и PROPERTY.

Кроме того, мы разметили датасет (BM), следуя BIO-схеме разметки последовательностей, и обучили на нем три NER модели – Distilroberta-base-mhg-charter-mlm, GHisBERT и SequenceTagger от Flair. Затем протестировали эти модели на текстах о Дитрихе, которые модель не видела. Модель

[5] https://github.com/Maryleya/NER-in-medieval-German-texts/blob/main/Dataset.csv

Flair показала лучшие результаты с F1-мерой, равной 0,74. Интересно, что это единственная модель из трех представленных, которая изначально была обучена на современном немецком языке. Такие невысокие результаты могут быть связаны с небольшим количеством данных для обучения, особенно это повлияло на тег GROUP, который не удалось выучить ни одной из моделей. Оставшиеся теги (PERSON и PLACE) в модели Flair достигли F1-меры, равной 0,75 и 0,72 соответственно.

Для дальнейших исследований по этой теме есть несколько направлений: увеличение количества данных для обучения за счет аннотирования большего числа текстов на средневерхненемецком языке, что потенциально приведет к повышению метрик качества, пополнение собранного датасета и добавление новых тегов в него, тестирование других NER моделей.

Литература

Aguilar 2022 – S. T. Aguilar. Multilingual Named Entity Recognition for Medieval Charters Using Stacked Embeddings and Bert-based Models. // *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages,* Marseille, France, June, 2022. P. 119–128.

Aguilar, Stutzmann 2021 – S. T. Aguilar, D. Stutzmann. Named Entity Recognition for French medieval charters. // *Proceedings of the Workshop on Natural Language Processing for Digital Humanities,* NIT Silchar, India, December, 2021. P. 37–46.

Akbik et al. 2018 – A. Akbik, D. A. J. Blythe, R. Vollgraf. Contextual String Embeddings for Sequence Labeling. // *International Conference on Computational Linguistics*, 2018.

Akbik et al. 2019 – A. Akbik, T. Bergmann, D. Blythe,K. Rasul, S. Schweter, R. Vollgraf. FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP. // *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, Minneapolis, Minnesota, June, 2019. P. 54–59.

Beck, Köllner 2023 – K. Beck, M. Köllner. GHisBERT – Training BERT from scratch for lexical semantic investigations across historical German language stages. // *Proceedings of the 4th Workshop on Computational Approaches to Historical Language Change*, Singapore, December, 2023. P. 33–45.

Besnier 2020 – C. Besnier. History to Myths: Social Network Analysis for Comparison of Stories over Time. // *LATECHCLFL*, 2020.

Besnier, Mattingly 2021 – C. Besnier, W. Mattingly. Named-Entity Dataset for Medieval Latin, Middle High German and Old Norse. // *Journal of Open Humanities Data*, 2021.

Bryn 2008 – S. Bryn. Creating Germany’s National Myth. The Nibelungenlied and its Homerian Context (Internet resource). https://brbl-archive.library.yale.edu/exhibitions/nibelungenlied/colophon.html. 2008.

Classen 2009 – A. Classen. The Nibelungenlied – Myth and History: A Middle High German Epic Poem at the Crossroads of Past and Present, Despair and Hope. // D. Konstan, K. A. Raaflaub (eds.). *Epic and History*. John Wiley & Sons, 2009. P. 262–279.

Curschmann 1987 – M. Curschmann. “Nibelungenlied” und “Klage”. // K. Ruh, G. Keil, W. Schröder (eds.). *Die deutsche Literatur des Mittelalters. Verfasserlexikon. Vol. 2.* Berlin, New York: Walter De Gruyter, 1987. P. 926–969.

Curschmann 1989 – M. Curschmann. Zur Wechselwirkung von Literatur und Sage: Das “Buch von Kriemhild” und “Das Buch von Bern”. // *Beiträge zur Geschichte der deutschen Sprache und Literatur* 111, 1989. P. 380–410.

Deutsches Textarchiv 2023 – Deutsches Textarchiv. Grundlage für ein Referenzkorpus der neuhochdeutschen Sprache (Internet resource). https://www.deutschestextarchiv.de/. 2023.

Eis 1953 – G. Eis. Zu Dietrichs Slawenkämpfen. // *Zeitschrift für deutsches Altertum und deutsche Literatur* 84, 1953. P. 7–84.

Fenniak 2024 – M. Fenniak. pypdf (Internet resource). *GitHub*. https://github.com/py-pdf/pypdf. 2024.

Gibbs, Johnson 1997 – M. E. Gibbs, S. M. Johnson. *Medieval German Literature: A Companion.* Routledge, 1997.

Haymes, Samples 1996 – E. R. Haymes, S. T. Samples. *Heroic legends of the North: an introduction to the Nibelung and Dietrich cycles*. New York: Garland, 1996.

Heinrichs 1955 – H. M. Heinrichs. Sivrit — Gernot — Kriemhilt. // *Zeitschrift Für Deutsches Altertum Und Deutsche Literatur* 86, 1955. P. 279–289.

Heinz 2015 – K. Heinz. *The International Centre for Archival Research (ICARUS)*. Ghent University, 2015.

Heinzle 1999 – J. Heinzle. *Einführung in die mittelhochdeutsche Dietrichepik.* Berlin, New York: De Gruyter, 1999.

Hoffmann 1974 – W. Hoffmann. *Mittelhochdeutsche Heldendichtung.* Berlin: Erich Schmidt, 1974.

Kenna, MacCarron 2017 – R. Kenna, P. MacCarron. A Networks Approach to Mythological Epics // R. Kenna, M. MacCarron, P. MacCarron (eds.). *Maths Meets Myths: Quantitative Approaches to Ancient Narratives.* Springer Verlag, 2017. P. 21–43.

Kragl 2015 – F. Kragl (ed.). *Der Wunderer*. Berlin, München, Boston: De Gruyter, 2015.

Lachmann 1878 – K. Lachmann. Der Nibelunge Noth und die Klage. *Nach der ältesten Überlieferung mit Bezeichnung des Unechten und mit den Abweichungen der gemeinen Lesart.* Reimer, 1878.

Layher 2009 – W. Layher. “She Was Completely Wicked”: Kriemhild as ‘Exemplum in a 13th-Century Sermon’. Image — Topos — Problem. // *Zeitschrift Für Deutsches Altertum Und Deutsche Literatur* 138, 2009. P. 344–360.

Lenter 2005 – A. Lenter. Namensverzeichnis ‹Rabenschlacht›. // E. Lienert, D. Wolter (eds.). *Rabenschlacht*. Berlin, Boston: Max Niemeyer Verlag, 2005. P. 237–264.

Lenter 2007 – A. Lenter. Namensverzeichnis ‹Alpharts Tod›. // E. Lienert, V. Meyer (eds.). *Alpharts Tod. Dietrich und Wenezlan*. Berlin, Boston: Max Niemeyer Verlag, 2007. P. 113–120.

Lenter, Wolter 2003 – A. Lenter, D. Wolter. Namensverzeichnis. ‹Dietrichs Flucht›. // E. Lienert, G. Beck (eds.). *Dietrichs Flucht*. Berlin, Boston: Max Niemeyer Verlag, 2003. P. 301–336.

Lienert 1999 – E. Lienert. Dietrich contra Nibelungen: Zur Intertextualität der historischen Dietrichepik. // *Beiträge zur Geschichte der deutschen Sprache und Literatur* 121. 1999. P. 23–46.

Lienert 2003 – E. Lienert. Rede und Schrift: Zur Inszenierung von Erzählen in mittelhochdeutscher Heldenepik. // C. Bertelsmeier-Kierst, C. Young (eds.). *Eine Epoche im Umbruch: Volkssprachige Literalität 1200-1300*. Tübingen: Niemeyer, 2003. P. 123–137.

Lienert 2015 – E. Lienert. *Mittelhochdeutsche Heldenepik*. Berlin: Erich Schmidt, 2015.

Lienert, Beck 2003 – E. Lienert, G. Beck (eds.). *Dietrichs Flucht*. Berlin, Boston: Max Niemeyer Verlag, 2003.

Lienter, Meyer 2007 – E. Lienert, V. Meyer (eds.). *Alpharts Tod. Dietrich und Wenezlan*. Berlin, Boston: Max Niemeyer Verlag, 2007.

Lienert, Wolter 2005 – E. Lienert, D. Wolter (eds.). *Rabenschlacht*. Berlin, Boston: Max Niemeyer Verlag, 2005.

Lienert et al. 2015 – E. Lienter, S. Kerth, S. Nierentz (eds.). *Rosengarten*. Berlin, München, Boston: De Gruyter, 2015.

McConnell 2002 – W. McConnell. Medieval German Heroic Epic. // F. G. Gentry (ed.). *A companion to middle high German literature to the 14th century*. Brill, 2002. P. 151–213.

McConnell et al. 2001 – W. McConnell, W. Wunderlich, F. Gentry, U. Mueller (eds.). *The Nibelungen Tradition: An Encyclopedia*. Routledge, 2001.

Millet 2008 – V. Millet. *Germanische Heldendichtung im Mittelalter*. Berlin, New York: de Gruyter, 2008.

Monasterium.net 2024 – Monasterium.net (Internet resource). http://monasterium.net:8181/mom/home. 2024.

Müller 2009 – J.-D. Müller (2009). *Das Nibelungenlied (3 ed.)*. Berlin: Erich Schmidt, 2009.

Nakayama 2018 – H. Nakayama. seqeval: A Python framework for sequence labeling evaluation (Internet resource). *Github*. https://github.com/chakki-works/seqeval. 2018.

Novotný et al. 2023 – V. Novotný, K. Luger, M. Štefánik, T. Vrabcová, A. Horák. People and Places of Historical Europe: Bootstrapping Annotation Pipeline and a New Corpus of Named Entities in Late Medieval Texts // *Findings of the Association for Computational Linguistics: ACL 2023,* Toronto, Canada, July, 2023. P. 14104–14113.

Piotrowski 2012 – *M. Piotrowski. Natural Language Processing for Historical Texts.* Springer Cham, 2012.

Reichert 2005 – H. Reichert (ed.). *Das Nibelungenlied. Nach der St. Galler Handschrift*. Berlin: de Gruyter, 2005.

Sahn et al. 2019 – V. Sanh, L. Debut, J. Chaumond, T. Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. // *ArXiv*, 2019.

Schramm 1965 – G. Schramm. Der Name Kriemhilt. // *Zeitschrift Für Deutsches Altertum Und Deutsche Literatur* 94, 1965. P. 39–57.

Schulz, Ketschik 2019 – S. Schulz, N. Ketschik. From 0 to 10 million annotated words: part-of-speech tagging for Middle High German. // *Language Resources and Evaluation* 53, 2019.

Schweter, Akbik 2021 – S. Schweter, A. Akbik. FLERT: Document-Level Features for Named Entity Recognition. // *ArXiv*, 2021.

Thuillard et al. 2018 – M. Thuillard, J.-L. Quellec, J. d'Huy. Computational Approaches to Myths Analysis: Application to the Cosmic Hunt. // *Nouvelle Mythologie comparée* 4, 2018. P. 1–32.

Valvekens 2023 – M. Valvekens. pdfminer.six (Internet resource). *GitHub*. https://github.com/pdfminer/pdfminer.six/tree/master. 2023.

Vollmer-Eicken 2007 – E. Vollmer-Eicken. Namensverzeichnis ‹Dietrich und Wenezlan›. // E. Lienert, V. Meyer (eds.). *Alpharts Tod. Dietrich und Wenezlan*. Berlin, Boston: Max Niemeyer Verlag, 2007. P. 121–122.

Voorwinden 2007 – N. Voorwinden. Dietrich von Bern: Germanic Hero or Medieval King? On the Sources of Dietrichs Flucht and Rabenschlacht. // *Neophilologus* 91, 2007. P. 243–259.

Zeige et al. 2022 – L. E. Zeige, G. Schnelle, M. Klotz, K. Donhauser, J. Gippert, R. Lühr. Deutsch Diachron Digital. Referenzkorpus Altdeutsch. Humboldt-Universität zu Berlin (Internet resource). http://www.deutschdiachrondigital.de/rea/. 2022.

Zeppezauer-Wachauer 2024 – K. Zeppezauer-Wachauer. Middle High German Conceptual Database | Mittelhochdeutsche Begriffsdatenbank (MHDBDB) (Internet resource). *GitHub*. https://github.com/Middle-High-German-Conceptual-Database. 2024.

Приложение

Ссылка на *GitHub*, где хранится собранный датасет, тетрадки с кодом и их описание: https://github.com/Maryleya/NER-in-medieval-German-texts