Федеральное государственное автономное образовательное
учреждение высшего образования
«Национальный исследовательский университет
«Высшая школа экономики»

Факультет компьютерных наук
Основная образовательная программа
Прикладная математика и информатика

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

Исследовательский проект на тему
"Атаки на мультиязычные модели"

Выполнил студент группы 171, 4 курса,
Биршерт Алексей Дмитриевич

Руководитель ВКР:
К. т. н., Доцент
Департамент больших данных и информационного поиска
Артемова Екатерина Леонидовна

Москва 2021

Содержание

Аннотация

Какие-то слова в абстракте. Какие-то слова в абстракте.
Ссылка на гитхаб с проектом - https://github.com/birshert/attack-lang-models.

Ключевые слова—Ключевые слова

Some words in abstract. Some words in abstract.
Github project link - https://github.com/birshert/attack-lang-models.

Keywords—Keywords

1 Введение

Последние несколько лет стали прорывными в области мультиязычных моделей и их обобщающей способности для других языков [1, 2, 7, 13]. Огромные мультиязычные модели выучивают универсальные языковые представления, что помогает им демонстрировать удивительные способности к переносу знаний с одного языка на другой. Простое дообучение предобученных моделей для какой-либо задачи на языке с большим количеством данных позволяет достичь хорошего качества на других языках.

Однако простой перенос между языками недостаточен для систем обработки естественного языка для понимания мультиязычных пользователей. Во многих сообществах в мире достаточно часто явление смешения кодов. Смешение кодов — это процесс, когда человек спонтанно смешивает различные языки внутри одного предложения или фразы. Такой феномен может проявляться как в письменной, так и в устной речи. Таким образом, важно сделать языковую модель устойчивой к смешению языков, чтобы модель адекватно работала со входными данными.

Несмотря на то, что реальные данные со смешением кодов очень важны для оценки качества языковых моделей, такие данные очень тяжело собирать и размечать в большом количестве.

В своей работе мы предполагаем, что качество моделей на адверсариальных атаках может служить нижней оценкой на реальное качество модели. Если языковая модель успешно справляется с адверсариальными пертурбациями со смешением кодов, то и в реальной жизни она будет успешно обрабатывать данные от мультиязычных пользователей.

В своей работе мы:

- Решаем задачу одновременного детектирования намерений пользователя и заполнения слотов для диалоговых помощников с помощью мультиязычных языковых моделей.

- Предлагаем две адверсариальные атаки по методу серого ящика — во

время атаки мы имеем доступ к ошибке модели на заданных данных. Насколько нам известно, это одни из первых мультиязычных адверсариальных атак для вышеописанной задачи.

- Предлагаем метод адверсариального предобучения.

В результате работы мы ожидаем получить следующие результаты:

- Мультиязычные модели обучены решать задачу заполнения слотов и классификации интентов.
- Проведены две адверсариальные атаки на каждую модель и замерено качество моделей на адверсариальных данных.
- Оценено влияние метода адверсариального предобучения на качество моделей на тестовой выборке и после адверсариальных атак.

Все свои эксперименты мы будем проводить с современными мультиязычными моделями - m-BERT [2] и XLM-RoBERTa [1]. В качестве датасета мы будем использовать корпус MultiAtis++ [14].

Актуальность темы подтверждается повышенным интересом со стороны научного сообщества. После начала работы над исследованием вышло как минимум три статьи на эту тему — две в марте [6, 10] и одна в конце апреля [9] 2021 года.

2 Обзор литературы

2.1 Мультиязычные модели

Языки с небольшим количеством данных часто не могут предоставить достаточного размера датасета для обучения с учителем. Существует подход для борьбы с этим, который заключается в построении кросс-язычных представлений. Эти представления нужно дообучать для специфичной задачи на

языке с большим количеством ресурсов, чтобы показывать хорошее качество на других, менее ресурсоёмких языках [5].

Вслед за успехом модели Трансформер [11], недавние мультиязычные модели такие как m-BERT [2] и XLM-RoBERTa [1] переносят парадигму «предобучение → дообучение под специфическую задачу» в мультиязычную область. Они предобучают энкодеры на основе архитектуры Трансформера на текстовых данных с различными задачами языкового моделирования. Затем эти предобученные энкодеры могут быть дообучены для конкретной задачи на ресурсоёмком языке для которого есть много размеченных данных. Это известно как кросс-язычный перенос знаний.

В одних недавних исследованиях кросс-язычного переноса знаний было показано, что качество модели на ранее не виденных тестовых языках сильно зависит от количества обучающих данных и размера контекста [7]. В [13] было показано, что m-BERT показывает очень сильную способность к кросс-язычному переносу знаний. m-BERT превосходит по качеству мультиязычные эмбеддинги в четырёх из пяти исследуемых задач без какой-либо информации о связи языков.

Более современная и более сложная модель XLM-RoBERTa [1] показывает лучшее, чем m-BERT качество, однако требует массивных объемов обучающих данных для хорошей работы. В своём исследовании авторы XLM-RoBERTa показывают, что их модель является самой сильной мультиязычной моделью на текущий момент.

m-BERT обучается на

2.2 Классификация интентов и заполнение слотов

[12]

2.3 Смешение кодов в адверсариальных атаках на мультиязычные модели

Основная - [10]. Побочная - [6]. Пуперпобочная - [9].

2.4 Машинный перевод и выравнивание слов

Перевод - [4].

Выравнивание - [3].

3 Основная часть

3.1 Обучение моделей на датасете MultiAtis++

В своей работе мы обучаем языковые модели решать задачу задачи одновременного детектирования намерений пользователя и заполнения слотов для диалоговых помощников, направленных на выполнение конкретной задачи. Эта задача заключается в классификации предложений и всех слов в предложении.

3.1.1 Датасет

В качестве датасета в своей работе мы выбрали датасет MultiAtis++ [14]. В этом датасете представлены семь языков из трёх языковых семей — Индо-Европейская (английский, немецкий, французский, испанский, португальский), Японо-рюкюская (японский) и Сино-тибетская (китайский). Датасет является параллельным корпусом для задачи классификации интентов и разметки слотов - в 2020 году он был переведён с английского языка на остальные шесть. В обучающей выборке содержится 4978 предложений для каждого языка, в тестовой 893 предложения для каждого языка.

Intent	atis_flight						
Utterance en	show	me	flights	from	montreal	to	orlando
Slot labels en	O	O	O	O	B-fromloc.city_name	O	B-toloc.city_name
Utterance de	Zeige	mir	Flüge	von	Montreal	nach	Orlando
Slot labels de	O	O	O	O	B-fromloc.city_name	O	B-toloc.city_name

Таблица 1: Пример объекта из датасета MultiAtis++. На примере представлен объект на английском и немецком языке.

Каждый объект в датасете состоит из предложения, меток слов в BIO формате и интента (Таблица (1)). Перед началом работы с датасетом мы произвели предварительную очистку — убрали из обучающей и тестовой выборок объекты, для которых на любом из семи языков количество слов и

количество слотов не совпадали. Таким образом, в обучающей выборке осталось 4884 объекта для каждого языка, в тестовой выборке 755 объектов для каждого языка. Для составления списка используемых слотов и интентов использовалась обучающая выборка на английском языке. Мы использовали 121 различную метку слотов и 23 различных метки интентов. Список id используемых объектов, а также списки используемых слотов и интентов можно найти в приложении.

3.1.2 Архитектура модели

В своей работе мы решаем задачу одновременной классификации интентов и разметки слотов в предложении с помощью одной модели. Модель имеет два выхода, первый предсказывает интенты, второй предсказывает метки слов. В качестве рассматриваемых архитектур были выбраны модели m-BERT [2] и XLM-RoBERTa [1]. Обе эти модели являются одними из самых сильных мультиязычных моделей на текущий момент. Каждая из них предобучена на более чем ста языках.

Обозначим количество блоков Трансформера за L, размер скрытых представлений за H и количество голов с внутренним вниманием за A. Тогда в используемой нами модели m-BERT L = 12, H = 768, A = 12, а суммарное количество параметров 110 миллионов. В используемой нами модели XLM-RoBERTa L = 12, H = 768, A = 12, а суммарное количество параметров 270 миллионов.

3.1.3 Обучение

В своей работе мы будем сравнивать модели, обученные на всей обучающей выборке и только на части обучающей выборки на английском языке. Таким образом мы сможем проверить насколько устойчивы к нашим атакам модели с разными вариантами обучения.

Введем краткие обозначения для удобства — модели XLM-RoBERTa будут обозначаться как «xlm-r», модели m-BERT будут обозначаться как «m-bert».

Если модель обучалась только на английской подвыборке, то мы будем добавлять в её название суффикс «en».

Каждая из моделей обучалась с одинаковыми гиперпараметрами - 10 эпох на обучающей выборке с длиной шага обучения 10^{-5} и размером батча в 64 объекта. В качестве функции ошибки использовалась кросс-энтропия:

$$L = -\frac{1}{n}\sum_{i=1}^{n}\left[y\log\left(\hat{y}\right)\right] \tag{1}$$

В своей работе мы будем использовать следующие метрики качества:

- Доля предложений, в которых правильно классифицирован интент:

$$\textbf{Intent accuracy} = \#\text{sentences}\left[\left(I_{pred} = I_{true}\right)\right] \tag{2}$$

- F1 мера для меток слотов (используется микро-усреднение по всем классам):

$$\textbf{Slots F1 score} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{3}$$

- Доля предложений, в которых правильно классифицирован интент и верно классифицированы все слоты:

$$\textbf{Semantic accuracy} = \#\text{sentences}\left[\left(I_{pred} = I_{true}\right) \wedge \left(S_{pred} = S_{true}\right)\right] \tag{4}$$

3.2 Адверсариальные атаки

В своей работе мы предлагаем два варианта gray-box адверсариальных атак — во время выполнения атаки мы имеем доступ к ошибке модели. Мы стремимся создать атаку такого рода, чтобы результирующая адверсариальная пертурбация предложения была как можно ближе к реалистичным предложениям со смешением кодов. Для этого мы заменяем часть токенов в предложении на их эквиваленты из других языков. Оценка качества на таких адверсариальных атаках может выступать в роли оценки снизу на качество

соответствующих моделей в аналогичных задачах при наличии реального смешения кодов во входных данных.

Так как большинство людей, которые могут использовать смешение кодов в своей речи, билингвы, то в основном смешение кодов происходит между парой языков [8]. Таким образом, в своей работе мы предлагаем анализировать атаки состоящие во встраивании одного языка в другой.

3.2.1 Общий вид атаки

Общий принцип атаки одинаковый для обоих предлагаемых вариантов. Разница между методами заключается в способе генерации кандидатов на замену токену на $i-$ой позиции. В своей работе мы предлагаем следующий вид атаки — пусть мы имеем целевую модель, пару пример-метка и встраиваемый язык (Алгоритм (1)). Тогда мы перебираем токены в предложении в случайном порядке и стремимся заменить токен на его эквивалент из встраиваемого языка. Если это приведёт к увеличению ошибки модели, то мы заменяем токен на предложенного кандидата.

Algorithm 1 Общая схема адверсариальной атаки

Require: Пара пример-метка x, y; целевая модель $\mathcal{M}$; встраиваемый язык $\mathbb{L}$
Ensure: Адверсариальный пример x'

```
L_x = GetLoss(M, x, y)
for i in permutation(len(x)) do
    Candidates = GetCandidates(M, x, y, token_id = i)
    Losses = GetLoss(M, Candidates)
    if Candidates and max(Losses) > L_x then
        L_x = max(Losses)
        x, y = Candidates[argmax(Losses)]
    end if
end for
    return x
```

3.2.2 Word level атака

Первый предлагаемый нами вариант атаки заключается в генерации эквивалентов из других языков с помощью перевода токенов на соответствующие

языки (Алгоритм (2)). Атакуя таким образом, мы строим грубую оценку снизу, так как при атаке мы не учитываем контекста предложений и не учитываем многозначность слов. Этот вариант схож с атакой PolyGloss [10]. Примеры атаки на тестовую выборку для модели XLM-RoBERTa можно найти в таблицах (2),(3) и (4).

Для перевода слов на другие языки мы используем модель машинного перевода M2M 100 от компании Facebook [4]. Она содержит 418 миллионов параметров.

Псевдокод функции ExtendSlotLabels можно найти в приложении (Алгоритм (5)).

Algorithm 2 Word-level атака

Require: Словарь переводов с исходного на встраиваемый язык $\mathbb{T}$
  function GETCANDIDATES($\mathcal{M}$, x, y, token_id)
    if x[token_id] in $\mathbb{T}[\mathbb{L}]$ **then**
      tokens = $\mathbb{T}[\mathbb{L}]$[x[token_id]]
      x[token_id] = tokens
      y[token_id] = ExtendSlotLabels(y[token_id], len(tokens))
    end if
    return x, y
  end function

Utterance en	which	flights	depart	from	philadelphia	and	arrive	in	atlanta	
Utterance adv	El	que	flights	Saída	from	philadelphia	and	arrive	in	Atlántico

Таблица 2: Пример 1 атаки модели XLM-RoBERTa (xlm-r) word-level атакой.

Utterance en	what	are	the	american	flights	from	newark	to	nashville
Utterance adv	what	are	El	american	flights	de	newark	to	Nashville

Таблица 3: Пример 2 атаки модели XLM-RoBERTa (xlm-r) word-level атакой.

Utterance en	list	flights	from	phoenix	arizona	to	ontario	california	wednesday	
Utterance adv	Lista	vuelos	from	El	Phoenix	arizona	para	El	Ontario	Californias	Miércoles

Таблица 4: Пример 3 атаки модели XLM-RoBERTa (xlm-r) word-level атакой.

3.2.3 Phrase-level атака

Второй предлагаемый нами вариант атаки заключается в генерации эквивалентов из других языков с помощью построения выравниваний между предложениями на разных языках. Одно предложение является переводом другого, для перевода можно использовать ту же модель машинного перевода [4], однако мы пользуемся тем, что у нас уже параллельный корпус. Кандидаты для каждого токена определяются как токены из предложения на встраиваемом языке, в которые был выровнен токен. Этот вариант атаки схож с атакой Bumblebee [10]. Примеры атаки на тестовую выборку для модели XLM-RoBERTa можно найти в таблицах (5),(6) и (7).

Для построения выравниваний мы используем модель awesome-align на основе m-BERT [3].

Algorithm 3 Phrase-level атака

Require: Выравнивание предложения на исходном языке к предложению на целевом языке $\mathbb{A}$

```
function GetCandidates(M, x, y, token_id)
    if x[token_id] in A[L] then
        tokens = A[L][x[token_id]]
        x[token_id] = tokens
        y[token_id] = ExtendSlotLabels(y[token_id], len(tokens))
    end if
    return x, y
end function
```

Utterance en	what	flights	travel	from	las	vegas	to	los	angeles
Utterance adv	what	flights	partem	from	Las	Vegas	to	Los	Angeles

Таблица 5: Пример 1 атаки модели XLM-RoBERTa (xlm-r) phrase-level атакой.

Utterance en	list	the	airlines	with	flights	to	or	from	denver
Utterance adv	list	the	airlines	with	flights	para	or	from	denver

Таблица 6: Пример 2 атаки модели XLM-RoBERTa (xlm-r) phrase-level атакой.

Utterance en	i want to fly from milwaukee to los angeles
Utterance adv	i quero to fly from milwaukee to Los angeles

Таблица 7: Пример 3 атаки модели XLM-RoBERTa (xlm-r) phrase-level атакой.

3.3 Метод адверсариального предобучения для защиты от адверсариальных атак

В своей работе мы предлагаем метод защиты от предложенных выше адверсариальных атак. Гипотеза заключается в том, что данный метод позволит увеличить качество не только на адверсариальных пертурбациях, но и на реальных данных со смешением кодов.

Предлагаемый нами метод адверсариального предобучения состоит из нескольких шагов:

1 Генерация выборки для задачи маскированного моделирования языка.

2 Дообучение тела мультиязычной модели на сгенерированной выборке в режиме предсказания маскированных токенов.

3 Загрузка дообученного тела модели перед началом обучения для задачи одновременного заполнения слотов и классификации интентов.

3.3.1 Генерация адверсариальной выборки

Для генерации выборки используется адаптация алгоритма phrase-level адверсариальной атаки (Алгоритм (4)). Разница заключается в том, что токены заменяются на их эквиваленты с некоторой вероятностью. Таким образом, для генерации выборки не требуется обученная модель.

Выборка является конкатенацией сгенерированных выборок для всех шести языков кроме английского представленных в датасете. Каждая из подвыборок генерируется встраиванием целевого языка в обучающую выборку да-

Algorithm 4 Генерация адверсариальной выборки

Require: Обучающая выборка датасета X, набор встраиваемых языков $\mathbb{L}_1, \ldots \mathbb{L}_n$

Ensure: Адверсариальная выборка X'

```
X' = [ ]
for 𝕃 in 𝕃_1, ... 𝕃_n do
    for x in X do
        for i in permutation(len(x)) do
            Candidates = GetCandidates(ℳ, x, y, token_id = i)
            if Candidates and 𝒰(0, 1) > 0.5 then
                x, _ = random.choice(Candidates)
            end if
        end for
        X'.append(x)
    end for
end for
return X'
```

тасета MultiAtis++ на английском языке. Псевдокод функции GetCandidates представлен в секции про атаки (Алгоритм (3)).

После генерации у нас получается 6 подвыборок по 4884 предложения в каждой. Итоговая выборка состоит из 29304 предложений, мы делим эту выборку в отношении 9 к 1 на обучающую и тестовую.

3.3.2 Дообучение тела модели

После генерации адверсариальной выборки мы дообучаем предобученную мультиязычную модель на этой выборке. Модель обучается в режиме задачи маскированного моделирования языка.

Для обучения модели для такой задачи мы отбираем 15% токенов и предсказываем их с помощью модели. 80% отобранных токенов заменяются на токен маски, 10% заменяются на случайные слова из словаря, остальные 10% остаются неизменными [2]. Мы дообучаем обе мультиязычные модели m-BERT и XLM-RoBERTa с одинаковыми гиперпараметрами - 10 эпох с размером батча 64 и длиной шага 10^{-5}. После дообучения мы сохраняем тело модели для дальнейшего использования.

3.3.3 Загрузка дообученного тела модели

Перед обучением мультиязычной модели для задачи одновременного заполнения слотов и классификации интентов мы загружаем дообученное тело модели.

Для моделей, которые были предобучены с помощью метода адверсариального предобучения, мы будем добавлять в название суффикс «adv» (3.1.3).

3.4 Результаты

3.4.1 Решение задачи классификации интентов и заполнения слотов

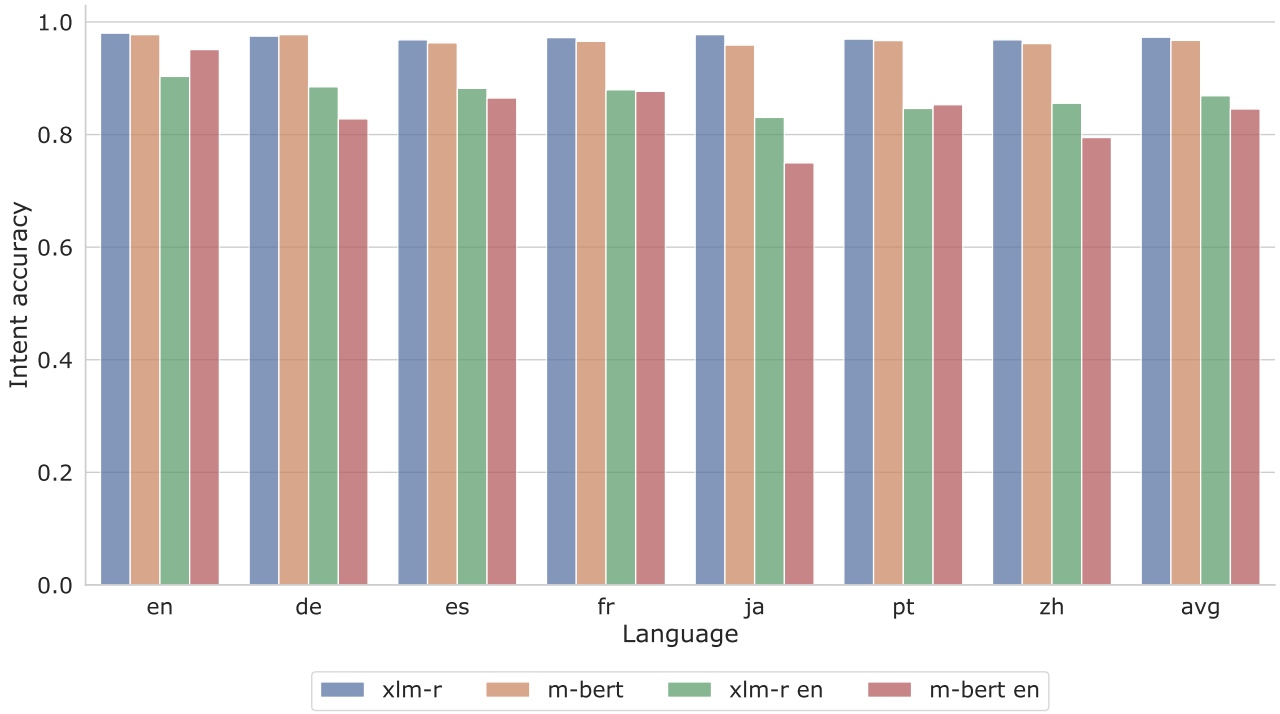

Рис. 1: Сравнение моделей между собой **на тестовой выборке** датасета MultiAtis++ по метрике **Intent accuracy**.

3.4.2 Качество моделей после адверсариальных атак

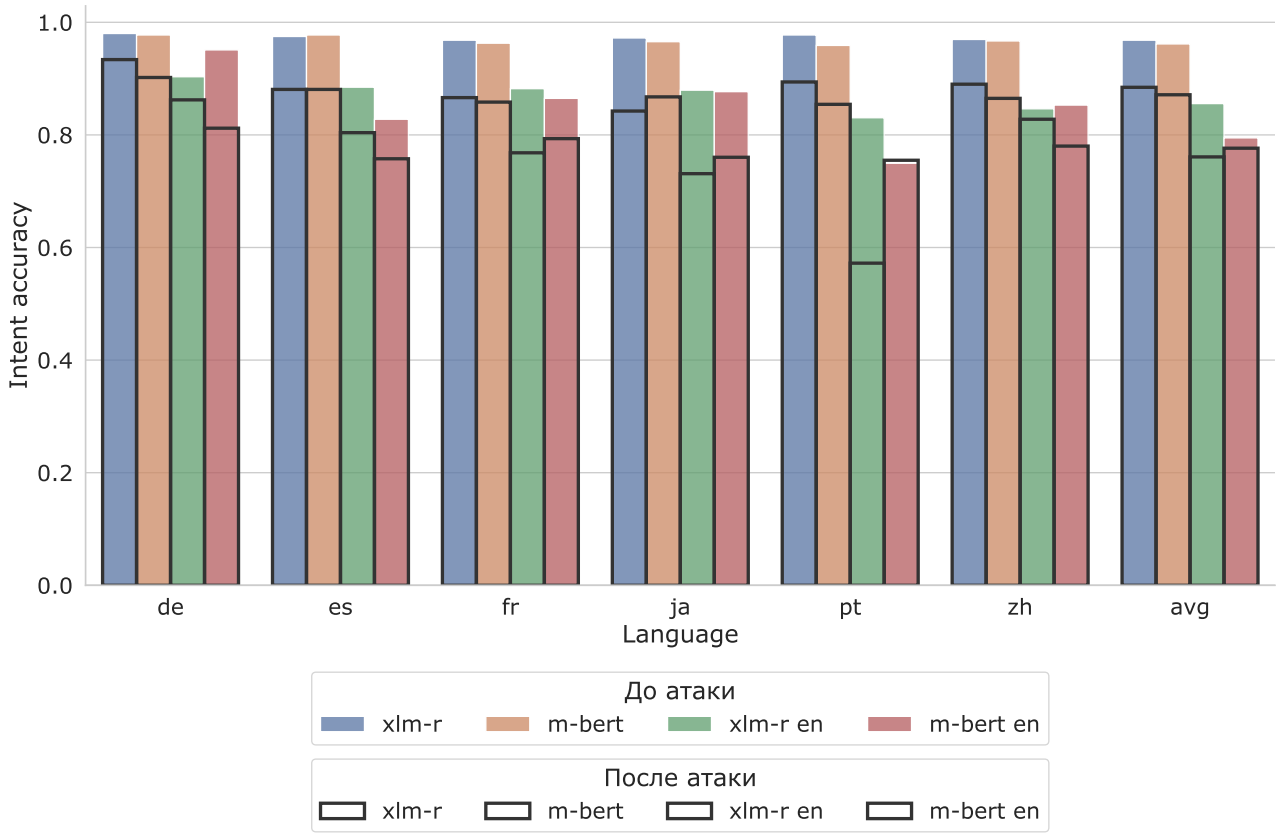

Рис. 2: Сравнение моделей между собой после **word-level** атаки на тестовую выборку датасета MultiAtis++ по метрике **Intent accuracy**.

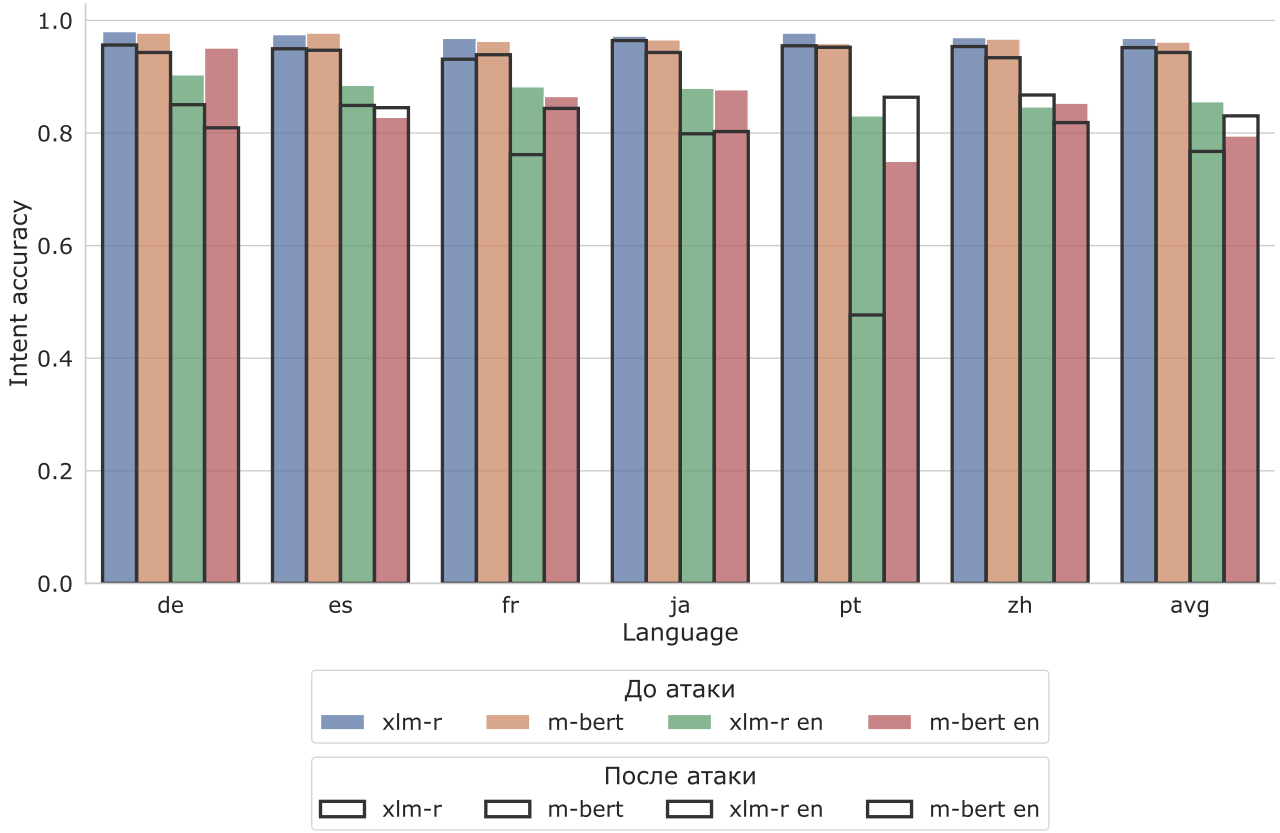

Рис. 3: Сравнение моделей между собой после **phrase-level** атаки на тестовую выборку датасета MultiAtis++ по метрике **Intent accuracy**.

3.4.3 Влияние метода адверсариального предобучения

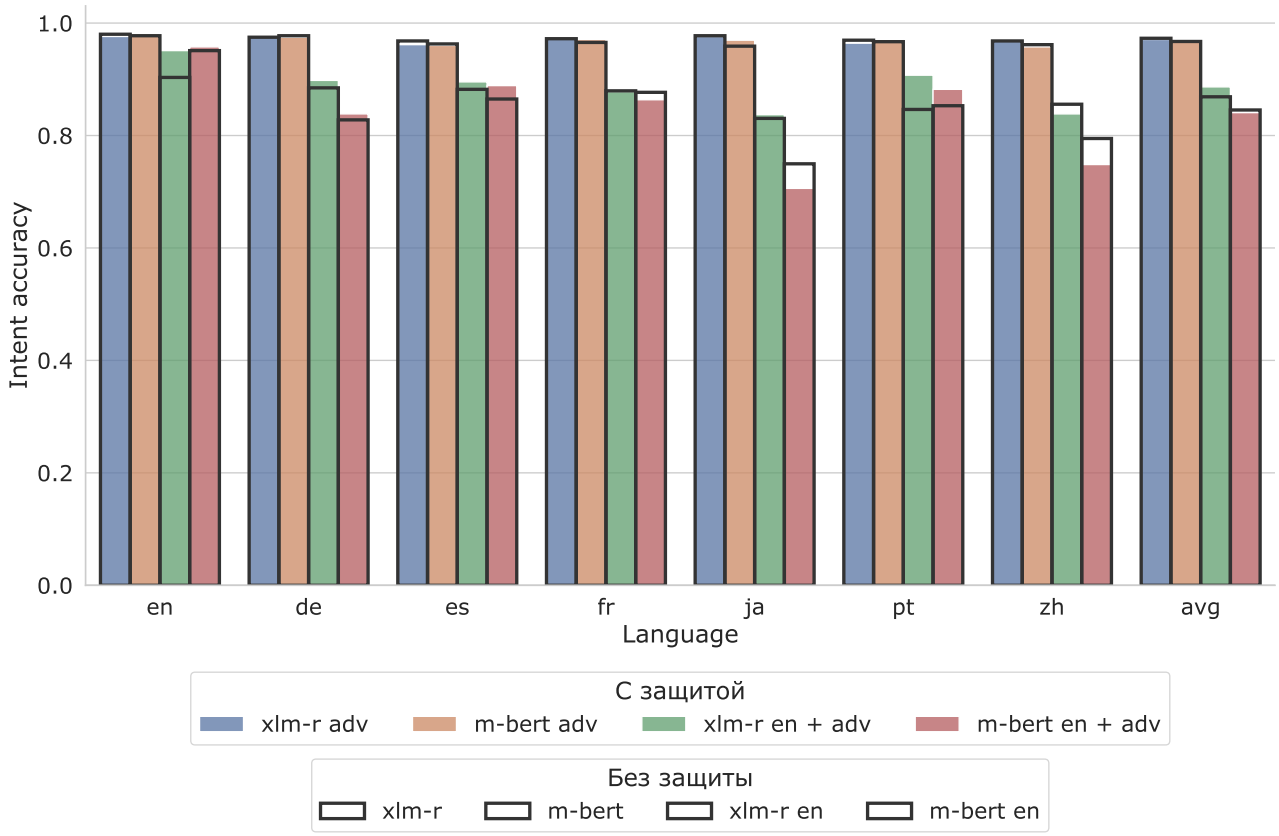

Рис. 4: Сравнение моделей **с защитой** между собой **на тестовой выборке** датасета MultiAtis++ по метрике **Intent accuracy**.

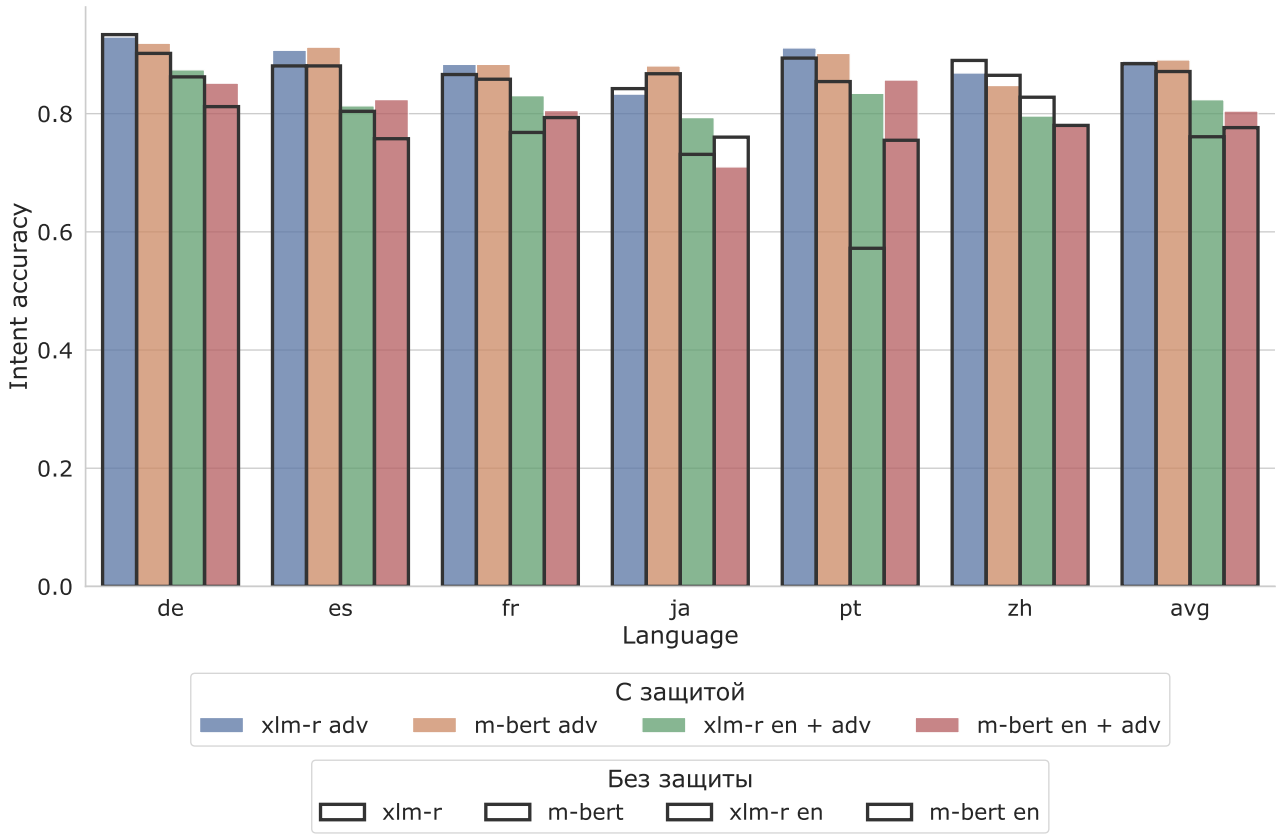

Рис. 5: Сравнение моделей **с защитой** между собой после **word-level** атаки на тестовую выборку датасета MultiAtis++ по метрике **Intent accuracy**.

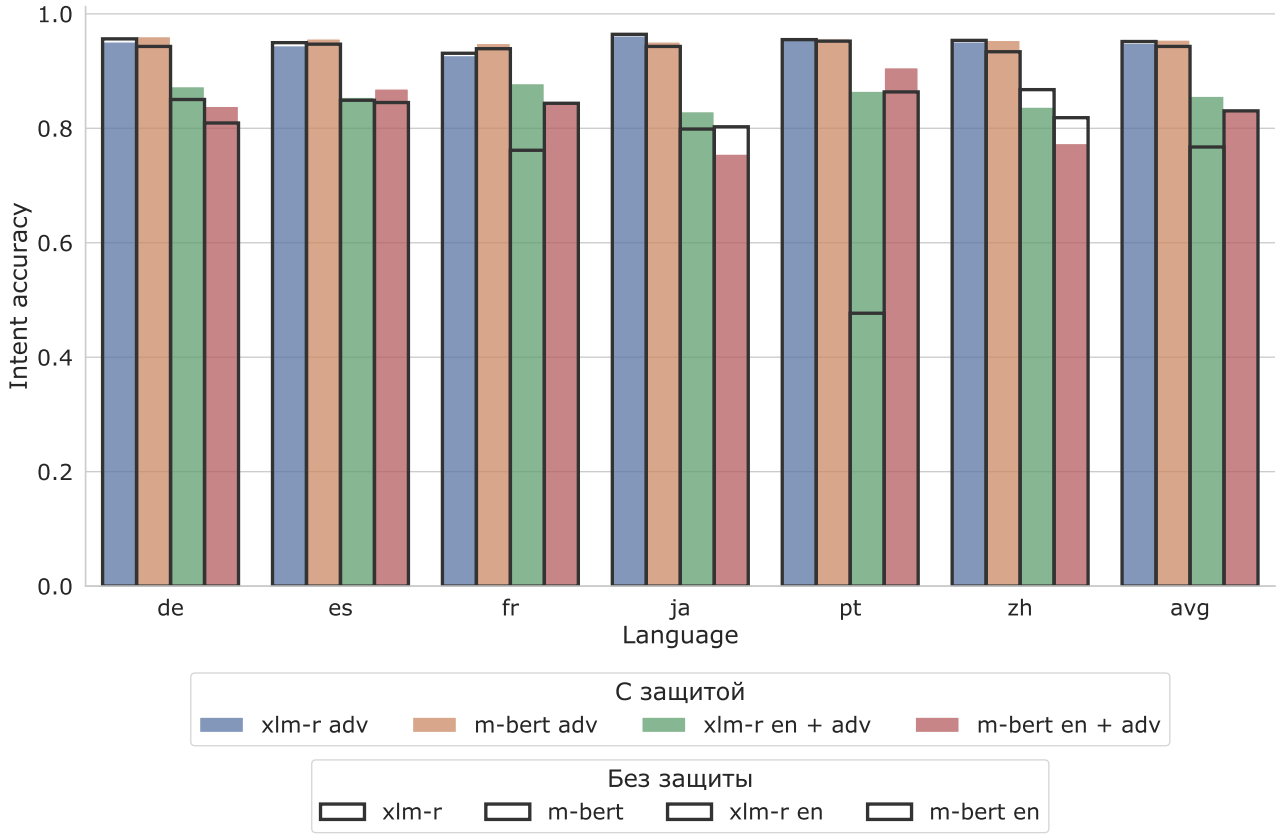

Рис. 6: Сравнение моделей **с защитой** между собой после **phrase-level** атаки на тестовую выборку датасета MultiAtis++ по метрике **Intent accuracy**.

4 Заключение

Список литературы

[1] Alexis Conneau и др. «Unsupervised Cross-lingual Representation Learning at Scale». В: *ACL*. 2020.

[2] Jacob Devlin и др. «BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding». В: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 2019, с. 4171—4186.

[3] Zi-Yi Dou и Graham Neubig. «Word Alignment by Fine-tuning Embeddings on Parallel Corpora». В: *EACL*. 2021.

[4] Angela Fan и др. «Beyond English-Centric Multilingual Machine Translation». В: *ArXiv* abs/2010.11125 (2020).

[5] Alexandre Klementiev, Ivan Titov и Binod Bhattarai. «Inducing Crosslingual Distributed Representations of Words». В: *Proceedings of COLING 2012*. Mumbai, India: The COLING 2012 Organizing Committee, дек. 2012, с. 1459—1474. URL: https://www.aclweb.org/anthology/C12-1089.

[6] Jitin Krishnan и др. «Multilingual Code-Switching for Zero-Shot Cross-Lingual Intent Prediction and Slot Filling». В: *ArXiv* abs/2103.07792 (2021).

[7] Chi-Liang Liu и др. «What makes multilingual BERT multilingual?» В: *ArXiv* abs/2010.10938 (2020).

[8] Shana Poplack, DAVID SANKOFF и CHRISTOPHER MILLER. «The social correlates and linguistic processes of lexical borrowing and assimilation». В: *Linguistics* 26 (1988), с. 47—104.

[9] Sebastin Santy, Anirudh Srinivasan и Monojit Choudhury. «BERTologiCoMix: How does Code-Mixing interact with Multilingual BERT?» В: *Proceedings of the Second Workshop on Domain Adaptation for NLP*. Kyiv, Ukraine: Association for Computational Linguistics, апр. 2021, с. 111—121. URL: https://www.aclweb.org/anthology/2021.adaptnlp-1.12.

[10] Samson Tan и Shafiq Joty. «Code-Mixing on Sesame Street: Dawn of the Adversarial Polyglots». В: *ArXiv* abs/2103.09593 (2021).

[11] Ashish Vaswani и др. «Attention is All you Need». В: *ArXiv* abs/1706.03762 (2017).

[12] H. Weld и др. «A survey of joint intent detection and slot-filling models in natural language understanding». В: *ArXiv* abs/2101.08091 (2021).

[13] Shijie Wu и Mark Dredze. «Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT». В: *EMNLP/IJCNLP*. 2019.

[14] Weijia Xu, Batool Haider и Saab Mansour. «End-to-End Slot Alignment and Recognition for Cross-Lingual NLU». В: *ArXiv* abs/2004.14353 (2020).

Приложения

A. Алгоритм замены слотов в атаке

Algorithm 5 Алгоритм замены слотов в атаке

```
function ExtendSlotLabels(slot_label, num_tokens)
    slot_labels = [slot_label]
    if num_tokens > 1 then
        if slot_label.startswith('B') then
            slot_labels += ['I' + slot_label[1:]] · (num_tokens - 1)
        else
            slot_labels ·= num_tokens
        end if
    end if
    return slot_labels
end function
```

Б. Примеры адверсариальных атак на модели

Utterance en	i need a daily flight from st. louis to milwaukee
Utterance adv	y need a Diariamente flight from El s. Luis para Milwaukee

Таблица 8: Пример атаки модели m-BERT (m-bert) word-level атакой.

Utterance en	i want a flight from toronto to san diego
Utterance adv	i want a voo from toronto para San diego

Таблица 9: Пример атаки модели m-BERT (m-bert) phrase-level атакой.

Utterance en	give me the flights from phoenix to milwaukee on american airlines
Utterance adv	give moi Le vols de phoenix à Milwaukee on american compagnies aériennes

Таблица 10: Пример атаки модели m-BERT (m-bert en) word-level атакой.

Utterance en	i'd like a morning flight from newark to los angeles
Utterance adv	gustaría me like un mañana flight from newark to los angeles

Таблица 11: Пример атаки модели m-BERT (m-bert en) phrase-level атакой.

Utterance en	show me flights from fort worth to san jose
Utterance adv	Show em me flights from fort Vale a pena Para san José

Таблица 12: Пример атаки модели m-BERT (m-bert adv) word-level атакой.

Utterance en	what are the flights from milwaukee to seattle
Utterance adv	what are Welche flights from milwaukee to seattle

Таблица 13: Пример атаки модели m-BERT (m-bert adv) phrase-level атакой.

Utterance en	give me the flights from phoenix to milwaukee on american airlines
Utterance adv	Geben Sie me the flights from Phoenix to Milwaukee on Amerikaner airlines

Таблица 14: Пример атаки модели m-BERT (m-bert en + adv) word-level атакой.

Utterance en	list the flights from indianapolis to memphis that leave before noon
Utterance adv	list the flights from Indianapolis to memphis that abheben before noon

Таблица 15: Пример атаки модели m-BERT (m-bert en + adv) phrase-level атакой.

Utterance en	how long does a flight from baltimore to san francisco take
Utterance adv	Cómo largo Se hace a vuelo de baltimore to san francisco Toma

Таблица 16: Пример атаки модели XLM-RoBERTa (xlm-r) word-level атакой.

Utterance en	list all flights from nashville to cleveland on sunday
Utterance adv	Lister tous vols from nashville to Cleveland on sunday

Таблица 17: Пример атаки модели XLM-RoBERTa (xlm-r) phrase-level атакой.

Utterance en	tell me the flights that leave philadelphia and go to dallas
Utterance adv	tell mich Die flights Das ist leave philadelphia und Gehen to dallas

Таблица 18: Пример атаки модели XLM-RoBERTa (xlm-r en) word-level атакой.

Utterance en	list flights from san jose to dallas on friday
Utterance adv	Lister flights de San José to dallas on vendredi Dallas

Таблица 19: Пример атаки модели XLM-RoBERTa (xlm-r en) phrase-level атакой.

Utterance en	give me the flights from pittsburgh to los angeles thursday evening
Utterance adv	de dar me El flights from El Pittsburgh to Los Ángeles Jueves por la noche

Таблица 20: Пример атаки модели XLM-RoBERTa (xlm-r adv) word-level атакой.

Utterance en	show me all delta airlines flights from montreal to orlando
Utterance adv	show me all delta airlines vuelos from montreal a orlando

Таблица 21: Пример атаки модели XLM-RoBERTa (xlm-r adv) phrase-level атакой.

Utterance en	list	flights	from	ontario	california	to	salt	lake	city	utah
Utterance adv	Lista	flights	from	ontario	Califórnia	to	Sal	Lago	Cidade	Utah

Таблица 22: Пример атаки модели XLM-RoBERTa (xlm-r en + adv) word-level атакой.

Utterance en	which	flights	go	from	new	york	to	miami	and	back
Utterance adv	which	Flüge	gehen	from	New	York	nach	Miami	and	zurück

Таблица 23: Пример атаки модели XLM-RoBERTa (xlm-r en + adv) phrase-level атакой.

B. Графики с результатами экспериментов

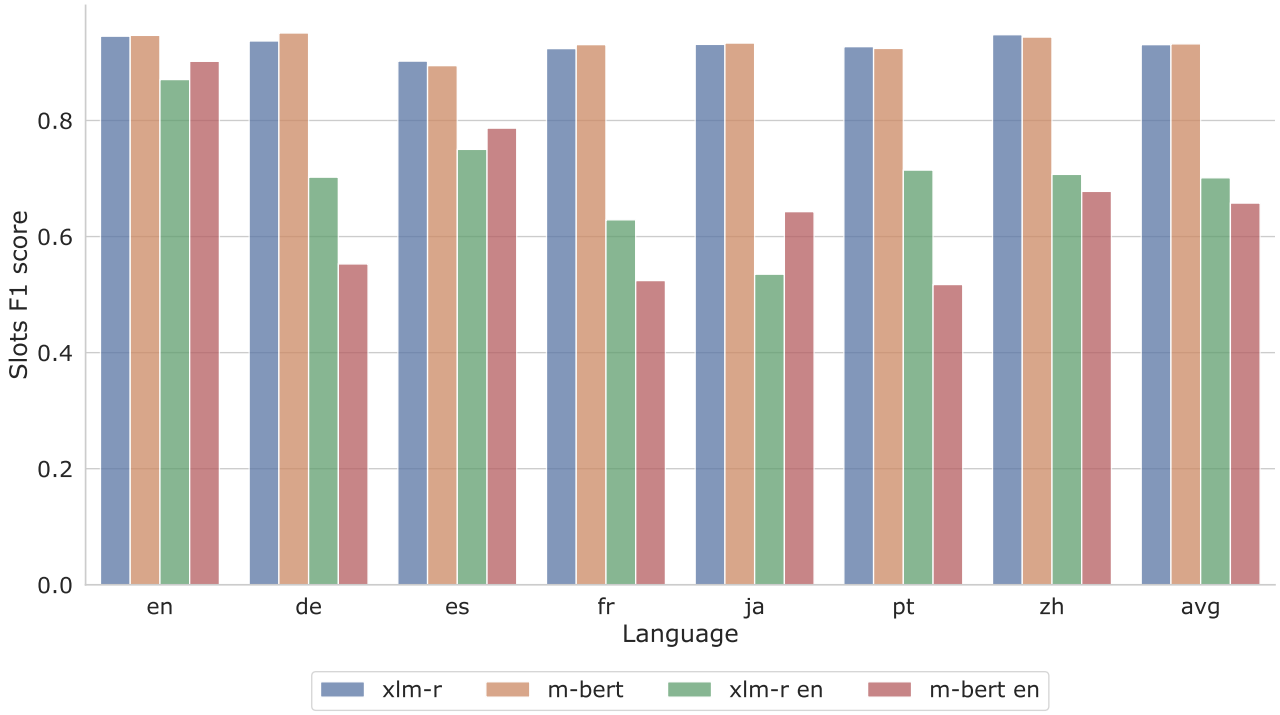

Рис. 7: Сравнение моделей между собой **на тестовой выборке** датасета MultiAtis++ по метрике **Slots F1 score**.

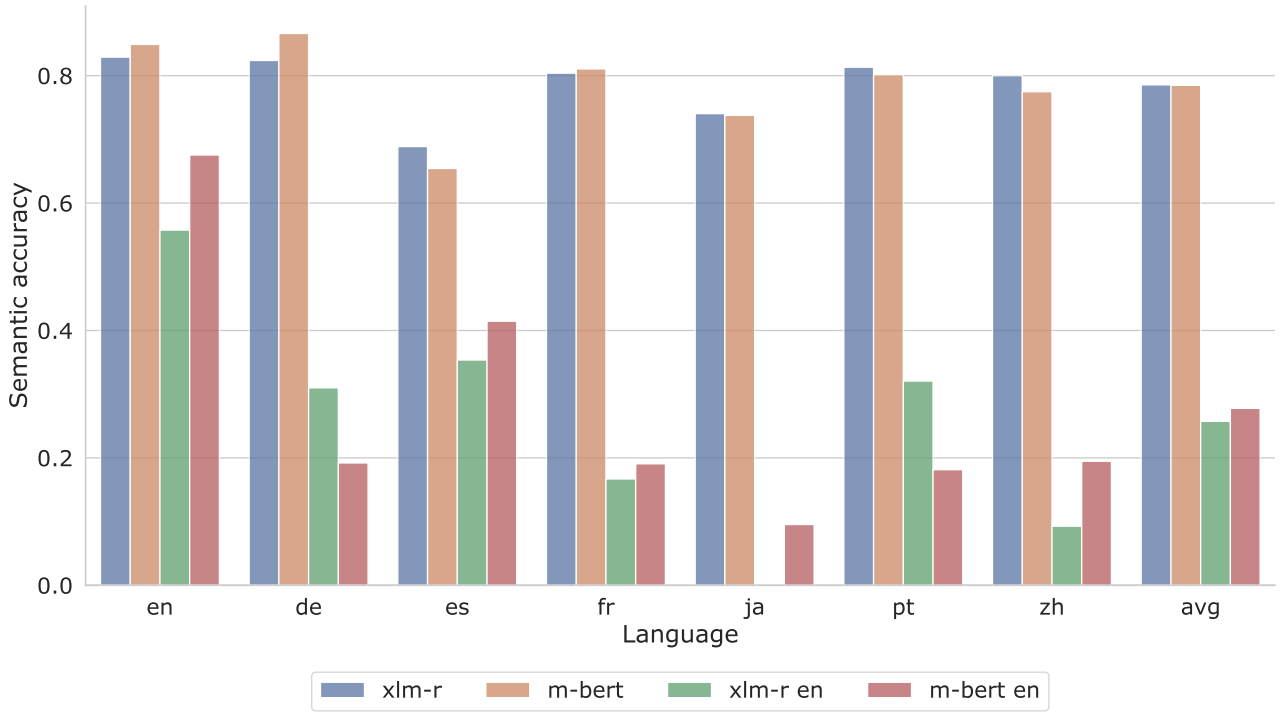

Рис. 8: Сравнение моделей между собой **на тестовой выборке** датасета MultiAtis++ по метрике **Semantic accuracy**.

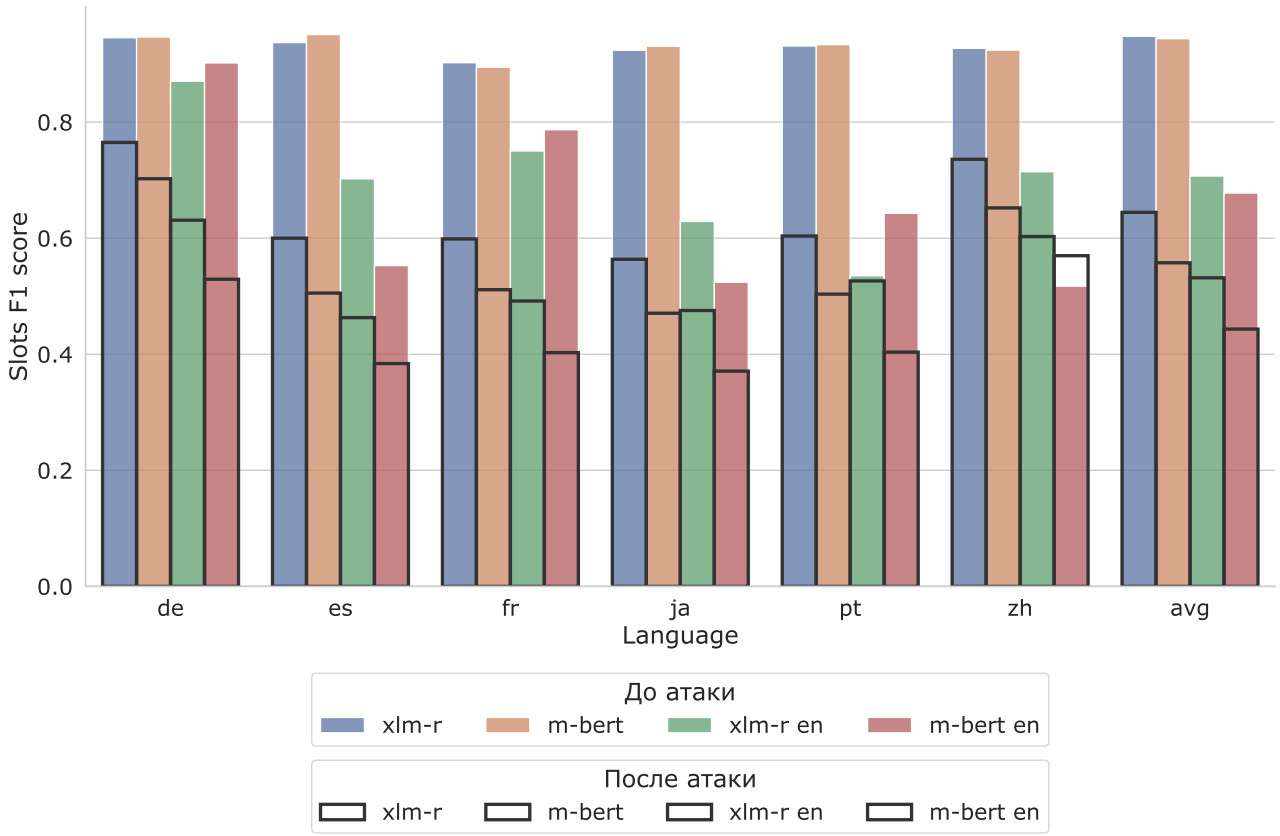

Рис. 9: Сравнение моделей между собой после **word-level** атаки на тестовую выборку датасета MultiAtis++ по метрике **Slots F1 score**.

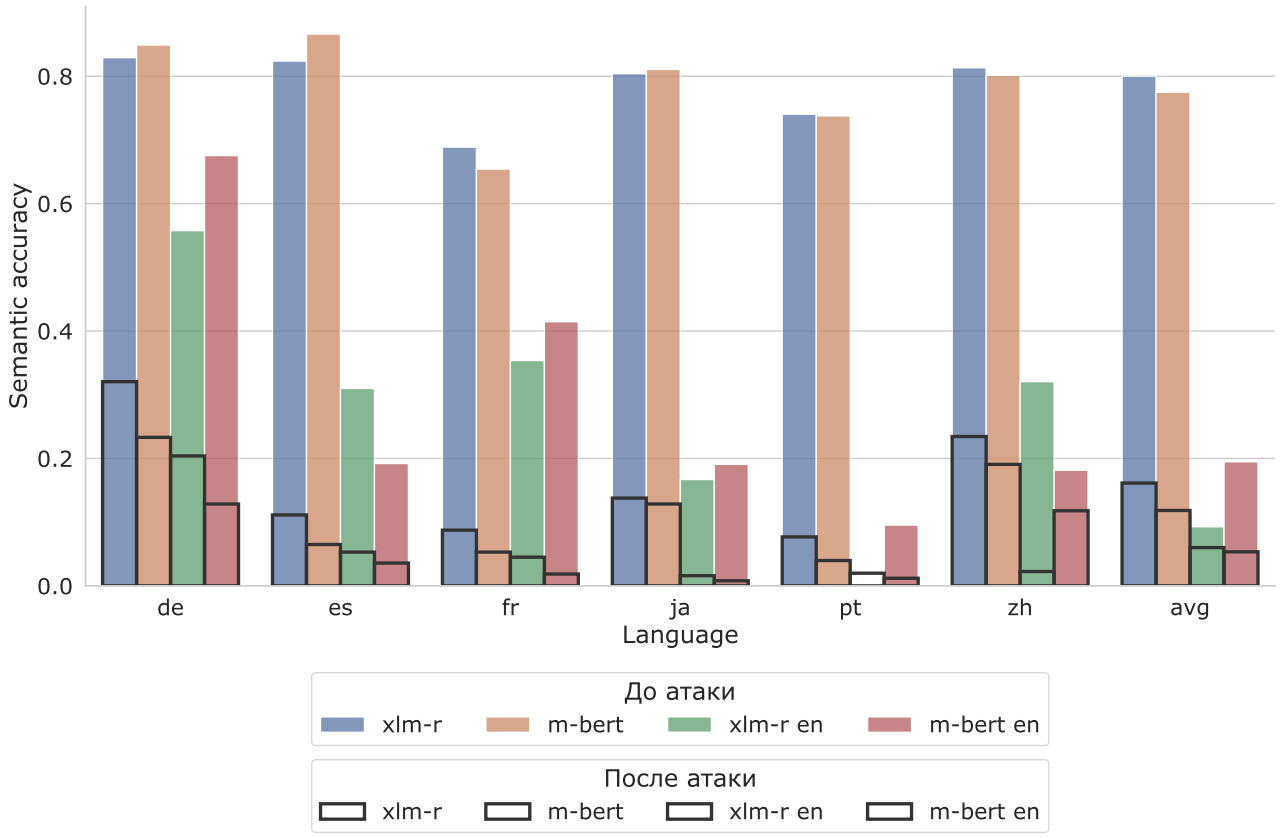

Рис. 10: Сравнение моделей между собой после **word-level** атаки на тестовую выборку датасета MultiAtis++ по метрике **Semantic accuracy**.

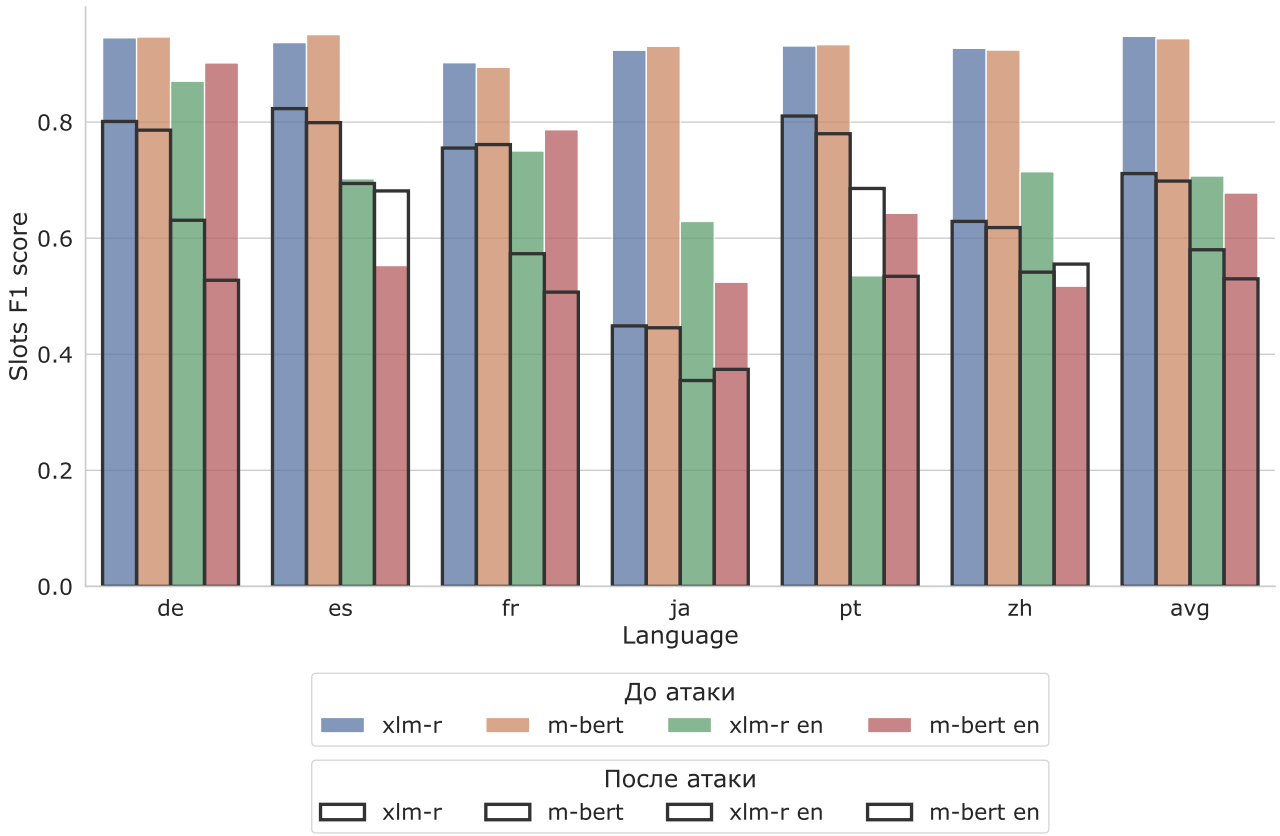

Рис. 11: Сравнение моделей между собой после **phrase-level** атаки на тестовую выборку датасета MultiAtis++ по метрике **Slots F1 score**.

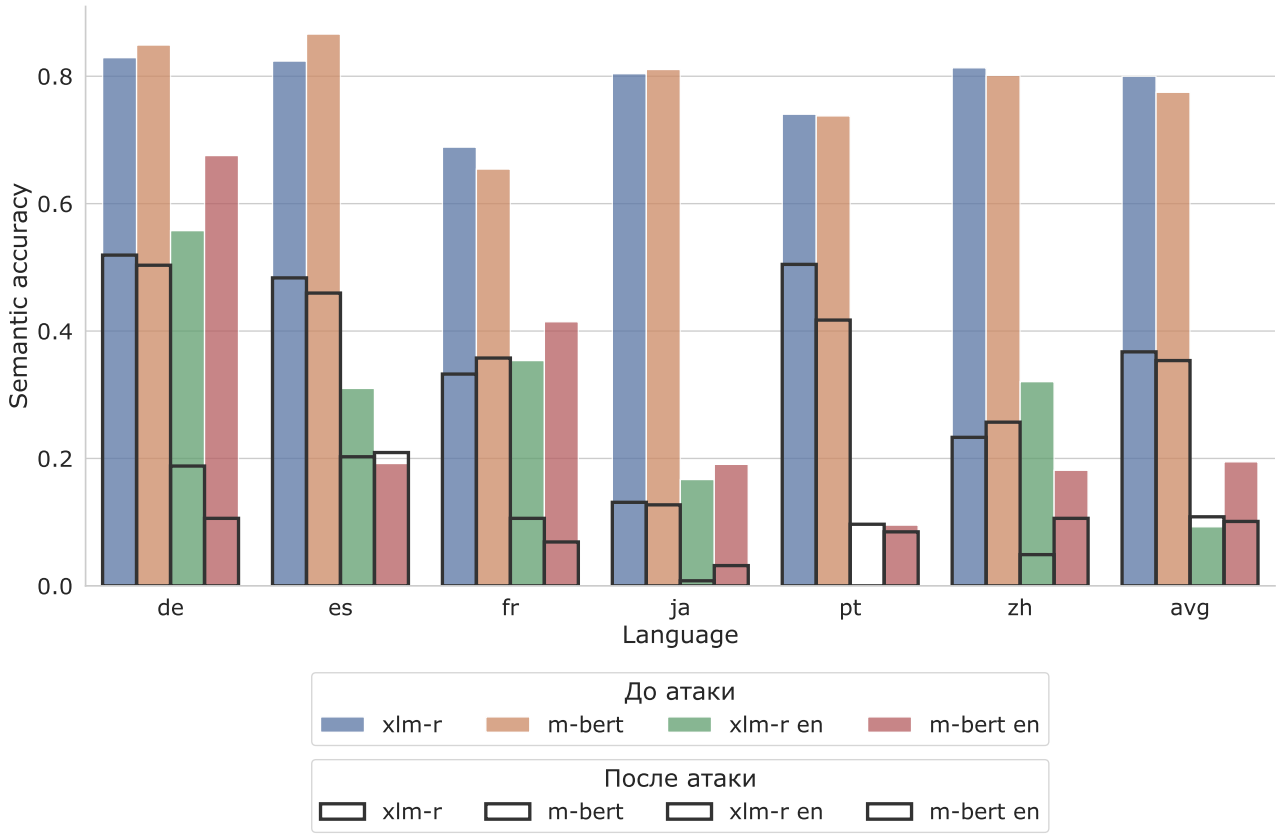

Рис. 12: Сравнение моделей между собой после **phrase-level** атаки на тестовую выборку датасета MultiAtis++ по метрике **Semantic accuracy**.

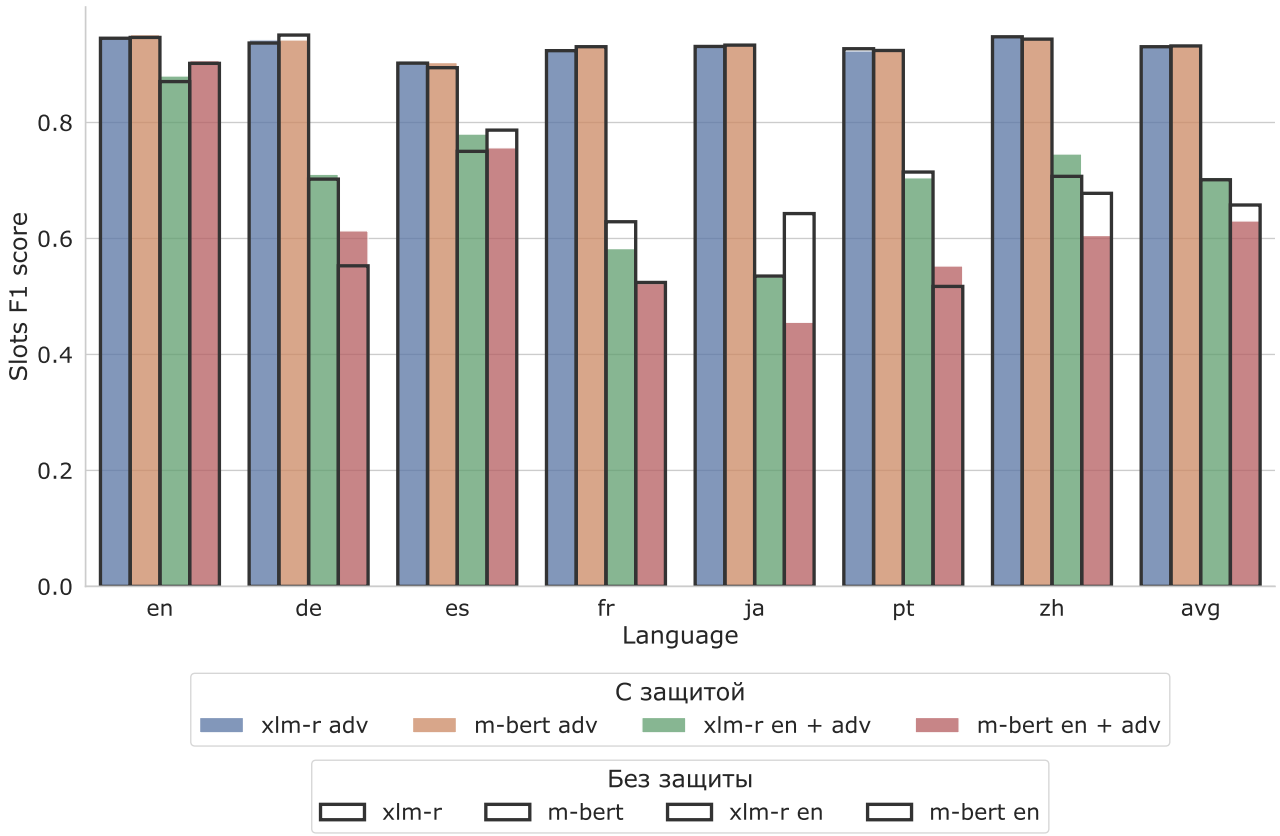

Рис. 13: Сравнение моделей **с защитой** между собой **на тестовой выборке** датасета MultiAtis++ по метрике **Slots F1 score**.

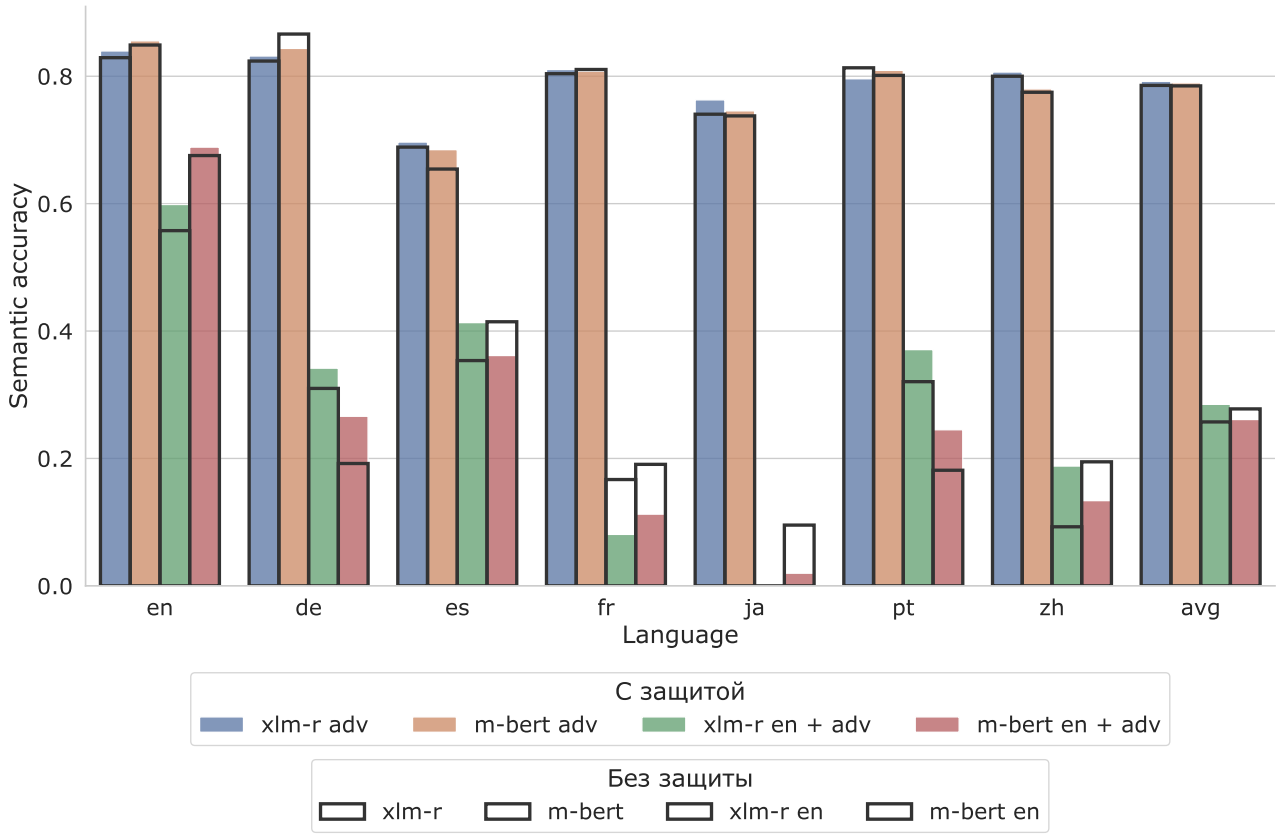

Рис. 14: Сравнение моделей **с защитой** между собой **на тестовой выборке** датасета MultiAtis++ по метрике **Semantic accuracy**.

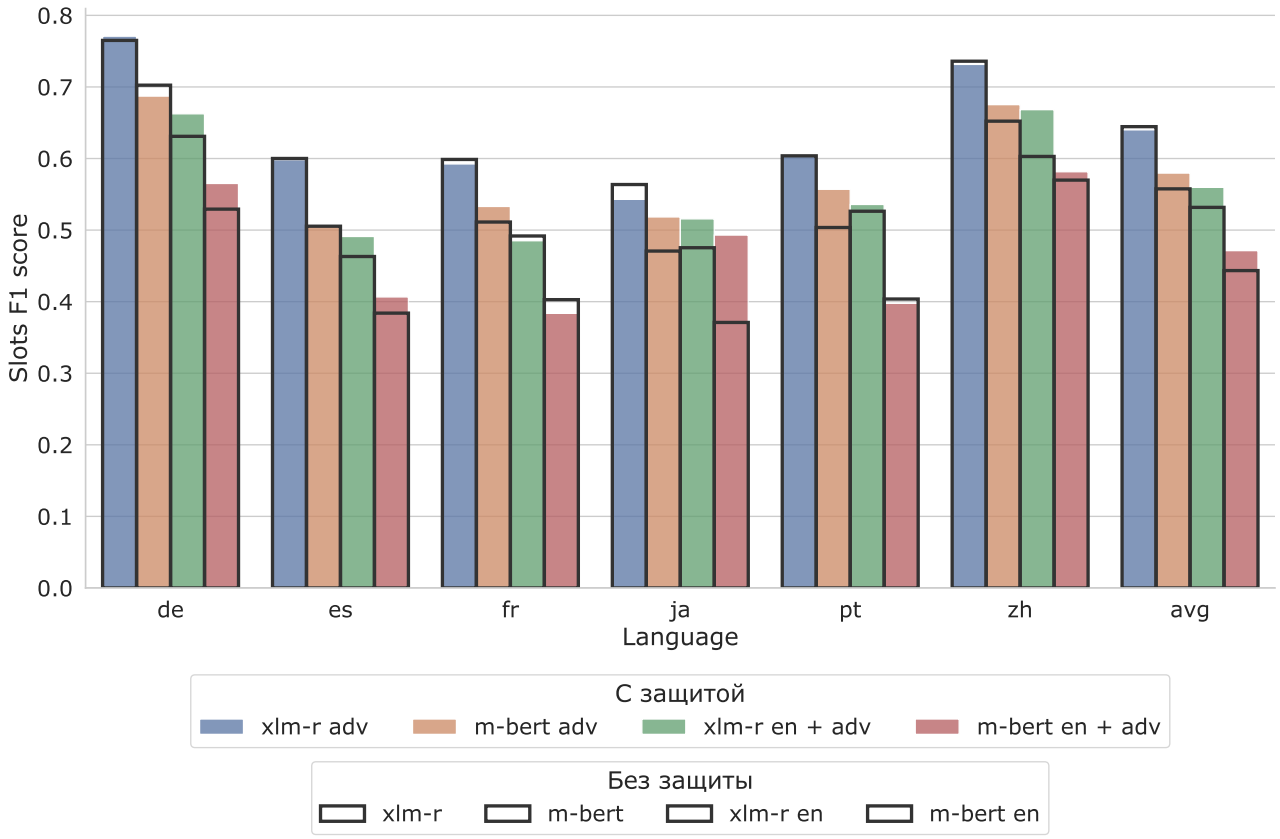

Рис. 15: Сравнение моделей **с защитой** между собой после **word-level** атаки на тестовую выборку датасета MultiAtis++ по метрике **Slots F1 score**.

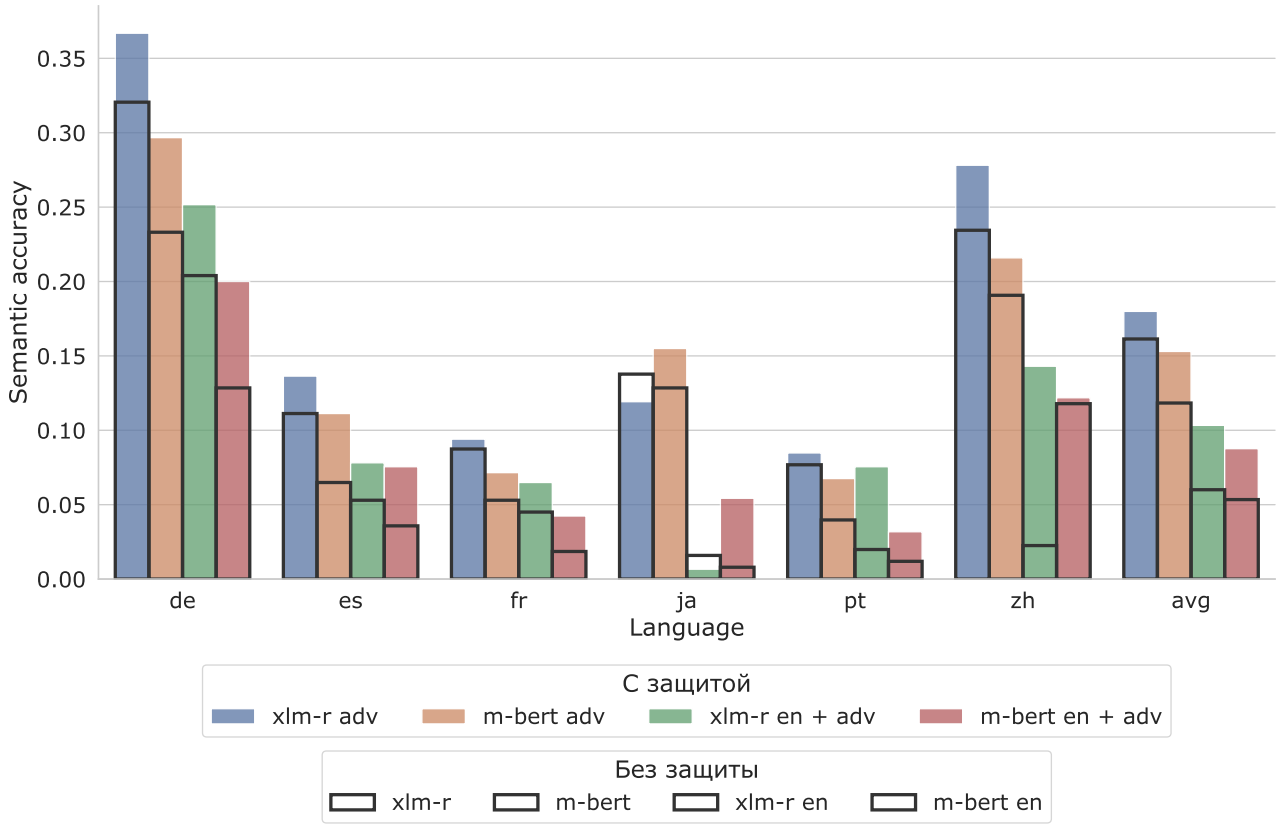

Рис. 16: Сравнение моделей **с защитой** между собой после **word-level** атаки на тестовую выборку датасета MultiAtis++ по метрике **Semantic accuracy**.

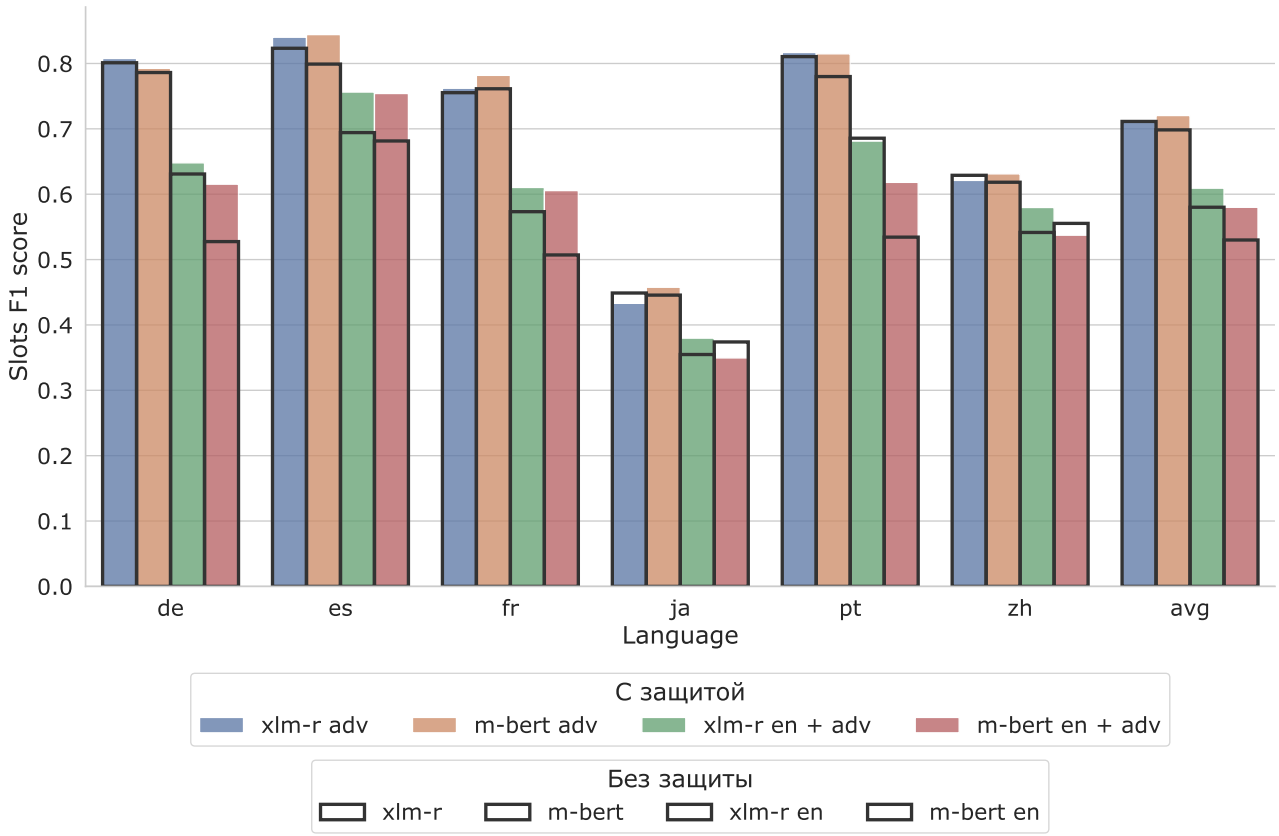

Рис. 17: Сравнение моделей **с защитой** между собой после **phrase-level** атаки на тестовую выборку датасета MultiAtis++ по метрике **Slots F1 score**.

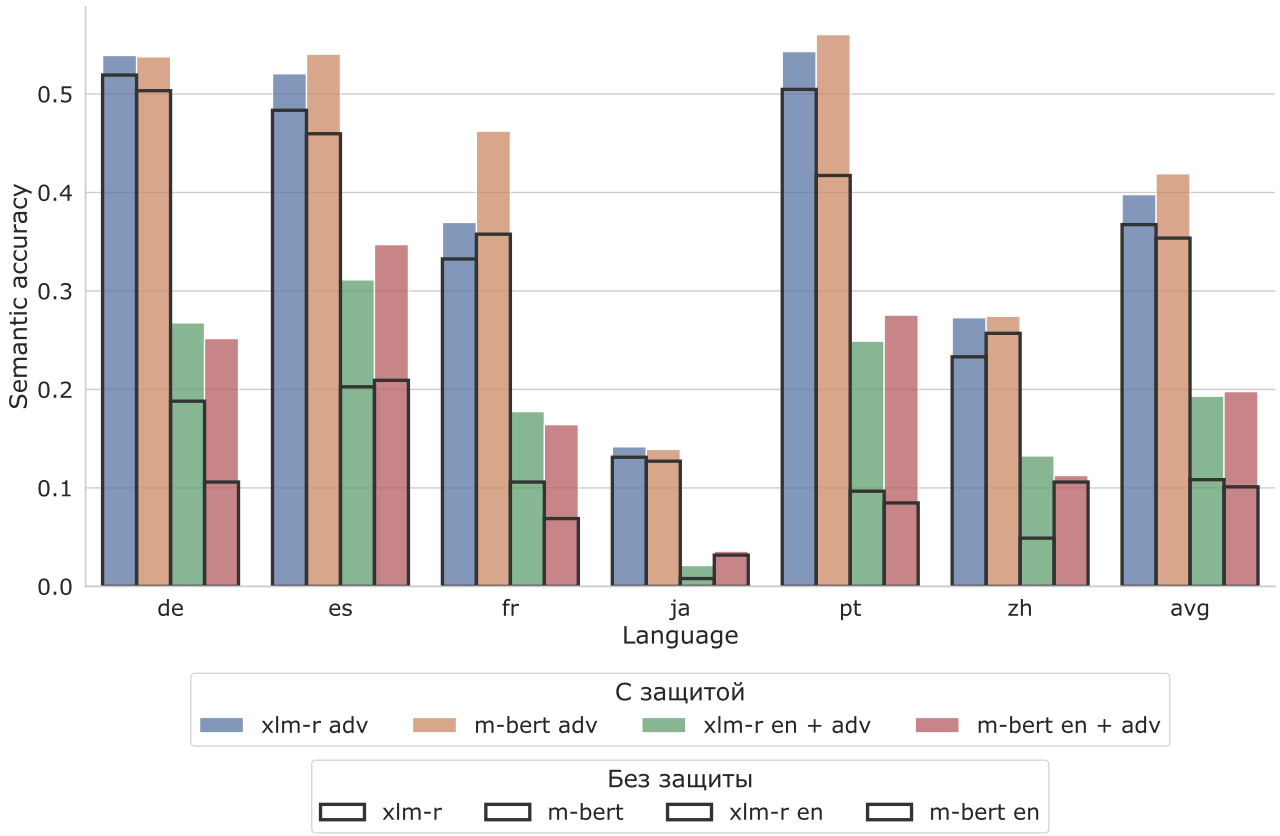

Рис. 18: Сравнение моделей **с защитой** между собой после **phrase-level** атаки на тестовую выборку датасета MultiAtis++ по метрике **Semantic accuracy**.

Г. Таблицы с результатами экспериментов

	en	de	es	fr	ja	pt	zh	avg
xlm-r	0.980	0.975	0.968	0.972	0.977	0.970	0.968	0.973
m-bert	0.977	0.977	0.963	0.966	0.959	0.967	0.962	0.967
xlm-r en	0.903	0.885	0.882	0.879	0.830	0.846	0.856	0.869
m-bert en	0.951	0.828	0.865	0.877	0.750	0.853	0.795	0.845

Таблица 24: Сравнение моделей между собой **на тестовой выборке** датасета MultiAtis++ по метрике **Intent accuracy**. По колонкам языки тестовых подвыборок, по рядам тестируемые модели.

	en	de	es	fr	ja	pt	zh	avg
xlm-r	0.945	0.937	0.902	0.924	0.931	0.927	0.948	0.931
m-bert	0.947	0.951	0.895	0.931	0.933	0.924	0.944	0.932
xlm-r en	0.871	0.702	0.750	0.629	0.535	0.715	0.707	0.701
m-bert en	0.902	0.553	0.787	0.524	0.643	0.517	0.678	0.658

Таблица 25: Сравнение моделей между собой **на тестовой выборке** датасета MultiAtis++ по метрике **Slots F1 score**. По колонкам языки тестовых подвыборок, по рядам тестируемые модели.

	en	de	es	fr	ja	pt	zh	avg
xlm-r	0.829	0.824	0.689	0.804	0.740	0.813	0.800	0.786
m-bert	0.849	0.866	0.654	0.811	0.738	0.801	0.775	0.785
xlm-r en	0.558	0.310	0.354	0.167	0.000	0.321	0.093	0.257
m-bert en	0.675	0.192	0.415	0.191	0.095	0.181	0.195	0.278

Таблица 26: Сравнение моделей между собой **на тестовой выборке** датасета MultiAtis++ по метрике **Semantic accuracy**. По колонкам языки тестовых подвыборок, по рядам тестируемые модели.

	de	es	fr	ja	pt	zh	avg
xlm-r	0.934	0.881	0.866	0.842	0.894	0.890	0.885
m-bert	0.902	0.881	0.858	0.868	0.854	0.865	0.871
xlm-r en	0.862	0.804	0.768	0.731	0.572	0.828	0.761
m-bert en	0.812	0.758	0.793	0.760	0.755	0.780	0.776

Таблица 27: Сравнение моделей между собой после **word-level** атаки на тестовую выборку датасета MultiAtis++ по метрике **Intent accuracy**. По колонкам встраиваемые языки, по рядам тестируемые модели.

	de	es	fr	ja	pt	zh	avg
xlm-r	0.765	0.600	0.599	0.564	0.604	0.736	0.645
m-bert	0.702	0.505	0.511	0.471	0.504	0.652	0.558
xlm-r en	0.631	0.463	0.492	0.475	0.526	0.603	0.532
m-bert en	0.529	0.384	0.403	0.371	0.404	0.570	0.443

Таблица 28: Сравнение моделей между собой после **word-level** атаки на тестовую выборку датасета MultiAtis++ по метрике **Slots F1 score**. По колонкам встраиваемые языки, по рядам тестируемые модели.

	de	es	fr	ja	pt	zh	avg
xlm-r	0.321	0.111	0.087	0.138	0.077	0.234	0.161
m-bert	0.233	0.065	0.053	0.128	0.040	0.191	0.118
xlm-r en	0.204	0.053	0.045	0.016	0.020	0.023	0.060
m-bert en	0.128	0.036	0.019	0.008	0.012	0.118	0.053

Таблица 29: Сравнение моделей между собой после **word-level** атаки на тестовую выборку датасета MultiAtis++ по метрике **Semantic accuracy**. По колонкам встраиваемые языки, по рядам тестируемые модели.

	de	es	fr	ja	pt	zh	avg
xlm-r	0.956	0.950	0.931	0.964	0.955	0.954	0.952
m-bert	0.943	0.947	0.939	0.943	0.952	0.934	0.943
xlm-r en	0.850	0.849	0.762	0.799	0.477	0.868	0.767
m-bert en	0.809	0.845	0.844	0.803	0.864	0.819	0.830

Таблица 30: Сравнение моделей между собой после **phrase-level** атаки на тестовую выборку датасета MultiAtis++ по метрике **Intent accuracy**. По колонкам встраиваемые языки, по рядам тестируемые модели.

	de	es	fr	ja	pt	zh	avg
xlm-r	0.801	0.823	0.755	0.449	0.810	0.629	0.711
m-bert	0.786	0.799	0.761	0.446	0.780	0.618	0.698
xlm-r en	0.631	0.694	0.573	0.355	0.686	0.541	0.580
m-bert en	0.528	0.681	0.507	0.374	0.534	0.555	0.530

Таблица 31: Сравнение моделей между собой после **phrase-level** атаки на тестовую выборку датасета MultiAtis++ по метрике **Slots F1 score**. По колонкам встраиваемые языки, по рядам тестируемые модели.

	de	es	fr	ja	pt	zh	avg
xlm-r	0.519	0.483	0.332	0.131	0.505	0.233	0.367
m-bert	0.503	0.460	0.358	0.127	0.417	0.257	0.354
xlm-r en	0.188	0.203	0.106	0.008	0.097	0.049	0.108
m-bert en	0.106	0.209	0.069	0.032	0.085	0.106	0.101

Таблица 32: Сравнение моделей между собой после **phrase-level** атаки на тестовую выборку датасета MultiAtis++ по метрике **Semantic accuracy**. По колонкам встраиваемые языки, по рядам тестируемые модели.

	en	de	es	fr	ja	pt	zh	avg
xlm-r adv	0.976	0.975	0.962	0.975	0.976	0.964	0.968	0.971
m-bert adv	0.981	0.975	0.960	0.971	0.970	0.971	0.958	0.969
xlm-r en + adv	0.951	0.898	0.895	0.878	0.837	0.907	0.838	0.886
m-bert en + adv	0.958	0.838	0.889	0.864	0.706	0.882	0.748	0.841

Таблица 33: Сравнение моделей **с защитой** между собой **на тестовой выборке** датасета MultiAtis++ по метрике **Intent accuracy**. По колонкам языки тестовых подвыборок, по рядам тестируемые модели.

	en	de	es	fr	ja	pt	zh	avg
xlm-r adv	0.948	0.942	0.906	0.927	0.933	0.924	0.950	0.933
m-bert adv	0.952	0.942	0.903	0.932	0.934	0.925	0.945	0.933
xlm-r en + adv	0.880	0.711	0.780	0.583	0.534	0.705	0.746	0.705
m-bert en + adv	0.907	0.613	0.756	0.522	0.456	0.553	0.605	0.630

Таблица 34: Сравнение моделей **с защитой** между собой **на тестовой выборке** датасета MultiAtis++ по метрике **Slots F1 score**. По колонкам языки тестовых подвыборок, по рядам тестируемые модели.

	en	de	es	fr	ja	pt	zh	avg
xlm-r adv	0.840	0.832	0.697	0.811	0.763	0.796	0.807	0.792
m-bert adv	0.856	0.844	0.685	0.808	0.746	0.809	0.780	0.790
xlm-r en + adv	0.599	0.342	0.413	0.081	0.001	0.371	0.188	0.285
m-bert en + adv	0.689	0.266	0.362	0.113	0.020	0.245	0.134	0.261

Таблица 35: Сравнение моделей **с защитой** между собой **на тестовой выборке** датасета MultiAtis++ по метрике **Semantic accuracy**. По колонкам языки тестовых подвыборок, по рядам тестируемые модели.

	de	es	fr	ja	pt	zh	avg
xlm-r adv	0.930	0.907	0.883	0.833	0.911	0.869	0.889
m-bert adv	0.919	0.913	0.883	0.881	0.902	0.848	0.891
xlm-r en + adv	0.874	0.813	0.830	0.793	0.834	0.796	0.824
m-bert en + adv	0.852	0.824	0.805	0.710	0.857	0.779	0.804

Таблица 36: Сравнение моделей **с защитой** между собой после **word-level** атаки на тестовую выборку датасета MultiAtis++ по метрике **Intent accuracy**. По колонкам встраиваемые языки, по рядам тестируемые модели.

	de	es	fr	ja	pt	zh	avg
xlm-r adv	0.771	0.598	0.592	0.543	0.604	0.731	0.640
m-bert adv	0.687	0.507	0.533	0.518	0.557	0.675	0.580
xlm-r en + adv	0.662	0.491	0.485	0.516	0.536	0.668	0.560
m-bert en + adv	0.565	0.407	0.384	0.493	0.398	0.582	0.471

Таблица 37: Сравнение моделей **с защитой** между собой после **word-level** атаки на тестовую выборку датасета MultiAtis++ по метрике **Slots F1 score**. По колонкам встраиваемые языки, по рядам тестируемые модели.

	de	es	fr	ja	pt	zh	avg
xlm-r adv	0.367	0.136	0.094	0.119	0.085	0.278	0.180
m-bert adv	0.297	0.111	0.072	0.155	0.068	0.216	0.153
xlm-r en + adv	0.252	0.078	0.065	0.007	0.075	0.143	0.103
m-bert en + adv	0.200	0.075	0.042	0.054	0.032	0.122	0.088

Таблица 38: Сравнение моделей **с защитой** между собой после **word-level** атаки на тестовую выборку датасета MultiAtis++ по метрике **Semantic accuracy**. По колонкам встраиваемые языки, по рядам тестируемые модели.

	de	es	fr	ja	pt	zh	avg
xlm-r adv	0.951	0.944	0.927	0.962	0.958	0.951	0.949
m-bert adv	0.960	0.956	0.948	0.951	0.956	0.954	0.954
xlm-r en + adv	0.873	0.854	0.878	0.829	0.865	0.837	0.856
m-bert en + adv	0.838	0.869	0.846	0.755	0.906	0.774	0.831

Таблица 39: Сравнение моделей **с защитой** между собой после **phrase-level** атаки на тестовую выборку датасета MultiAtis++ по метрике **Intent accuracy**. По колонкам встраиваемые языки, по рядам тестируемые модели.

	de	es	fr	ja	pt	zh	avg
xlm-r adv	0.808	0.840	0.762	0.433	0.817	0.621	0.713
m-bert adv	0.793	0.844	0.782	0.458	0.815	0.631	0.720
xlm-r en + adv	0.648	0.756	0.610	0.380	0.681	0.580	0.609
m-bert en + adv	0.615	0.754	0.606	0.350	0.618	0.537	0.580

Таблица 40: Сравнение моделей **с защитой** между собой после **phrase-level** атаки на тестовую выборку датасета MultiAtis++ по метрике **Slots F1 score**. По колонкам встраиваемые языки, по рядам тестируемые модели.

	de	es	fr	ja	pt	zh	avg
xlm-r adv	0.539	0.521	0.370	0.142	0.543	0.273	0.398
m-bert adv	0.538	0.540	0.462	0.139	0.560	0.274	0.419
xlm-r en + adv	0.268	0.311	0.177	0.021	0.249	0.132	0.193
m-bert en + adv	0.252	0.347	0.164	0.036	0.275	0.113	0.198

Таблица 41: Сравнение моделей **с защитой** между собой после **phrase-level** атаки на тестовую выборку датасета MultiAtis++ по метрике **Semantic accuracy**. По колонкам встраиваемые языки, по рядам тестируемые модели.