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

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

1.1 Описание предметной области

1.2 Актуальность работы

1.3 Цель и задачи работы

1.4 Постановка задачи

1.5 Ожидаемые результаты

1.6 Структура работы

2 Обзор литературы

2.1 Что-то первое

2.2 Что-то второе

2.3 Что-то третье

2.4 Что-то четвертое

3 Основная часть

[?]

	xlm-r	xlm-r en	xlm-r adv	xlm-r en + adv
Intent accuracy	0.980	0.902	0.980	0.963
Slot F1 score	0.944	0.870	0.948	0.899
Semantic accuracy	0.826	0.559	0.842	0.670
Loss	0.317	0.729	0.293	0.575

Таблица 1: Таблица сравнения моделей XLM-R между собой на тестовой выборке

	m-bert	m-bert en	m-bert adv	m-bert en + adv
Intent accuracy	0.979	0.952	0.975	0.948
Slot F1 score	0.947	0.899	0.952	0.908
Semantic accuracy	0.854	0.672	0.846	0.690
Loss	0.353	0.584	0.328	0.577

Таблица 2: Таблица сравнения моделей M-BERT между собой на тестовой выборке

	xlm-r	xlm-r en	xlm-r adv	xlm-r en + adv
Intent accuracy	0.885 ± 0.035	0.727 ± 0.081	0.893 ± 0.037	0.851 ± 0.035
Slot F1 score	0.642 ± 0.080	0.550 ± 0.069	0.651 ± 0.078	0.568 ± 0.065
Semantic accuracy	0.179 ± 0.097	0.065 ± 0.059	0.191 ± 0.105	0.089 ± 0.067
Loss	2.627 ± 0.727	3.232 ± 0.809	2.424 ± 0.667	2.624 ± 0.612

Таблица 3: Таблица сравнения моделей XLM-R после атаки Word level

	m-bert	m-bert en	m-bert adv	m-bert en + adv
Intent accuracy	0.866 ± 0.028	0.771 ± 0.032	0.863 ± 0.023	0.781 ± 0.046
Slot F1 score	0.556 ± 0.095	0.444 ± 0.083	0.585 ± 0.086	0.489 ± 0.064
Semantic accuracy	0.120 ± 0.079	0.056 ± 0.053	0.145 ± 0.088	0.090 ± 0.065
Loss	3.137 ± 0.701	3.335 ± 0.662	2.878 ± 0.611	3.019 ± 0.512

Таблица 4: Таблица сравнения моделей M-BERT после атаки Word level

	xlm-r	xlm-r en	xlm-r adv	xlm-r en + adv
Intent accuracy	0.947 ± 0.006	0.728 ± 0.136	0.954 ± 0.009	0.864 ± 0.040
Slot F1 score	0.708 ± 0.140	0.581 ± 0.109	0.721 ± 0.148	0.641 ± 0.129
Semantic accuracy	0.366 ± 0.156	0.105 ± 0.074	0.405 ± 0.164	0.228 ± 0.138
Loss	2.026 ± 1.152	2.860 ± 0.826	1.992 ± 1.248	1.943 ± 0.743

Таблица 5: Таблица сравнения моделей XLM-R после атаки Alignments

	m-bert	m-bert en	m-bert adv	m-bert en + adv
Intent accuracy	0.942 ± 0.004	0.828 ± 0.020	0.950 ± 0.005	0.818 ± 0.035
Slot F1 score	0.700 ± 0.127	0.536 ± 0.096	0.728 ± 0.137	0.577 ± 0.150
Semantic accuracy	0.348 ± 0.127	0.113 ± 0.055	0.406 ± 0.158	0.198 ± 0.113
Loss	2.118 ± 1.143	2.474 ± 0.591	1.935 ± 1.135	2.252 ± 0.825

Таблица 6: Таблица сравнения моделей M-BERT после атаки Alignments

4 Заключение

AAAAAAAAAAAAAAAAAA FUCK ME