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

[1]

3 Основная часть

	xlm-r	xlm-r adv pretrained	xlm-r pretrained	m-bert	m-bert adv pretrained	m-bert pretrained
intent_acc	0.943 ± 0.016	0.945 ± 0.020	0.970 ± 0.007	0.954 ± 0.010	0.960 ± 0.007	0.971 ± 0.008
slot_f1	0.898 ± 0.035	0.903 ± 0.043	0.937 ± 0.016	0.895 ± 0.040	0.898 ± 0.041	0.933 ± 0.021
sementic_frame_acc	0.669 ± 0.098	0.687 ± 0.109	0.815 ± 0.054	0.660 ± 0.104	0.684 ± 0.104	0.801 ± 0.066
loss	0.619 ± 0.131	0.589 ± 0.154	0.395 ± 0.066	0.518 ± 0.122	0.509 ± 0.127	0.397 ± 0.068

Таблица 1: Таблица сравнения моделей между собой на тестовой выборке

	en	de	es	fr	ja	pt	zh
intent_acc	0.971 ± 0.004	0.964 ± 0.008	0.947 ± 0.020	0.967 ± 0.007	0.951 ± 0.016	0.946 ± 0.020	0.954 ± 0.012
slot_f1	0.940 ± 0.010	0.930 ± 0.013	0.839 ± 0.041	0.903 ± 0.023	0.922 ± 0.010	0.905 ± 0.019	0.937 ± 0.011
sementic_frame_acc	0.803 ± 0.043	0.798 ± 0.047	0.521 ± 0.115	0.742 ± 0.063	0.693 ± 0.051	0.736 ± 0.060	0.742 ± 0.071
loss	0.384 ± 0.045	0.438 ± 0.071	0.731 ± 0.142	0.433 ± 0.070	0.535 ± 0.095	0.529 ± 0.136	0.481 ± 0.092

Таблица 2: Таблица сравнения качества для языков на тестовой выборке

	en	de	es	fr	ja	pt	zh
intent_acc	0.818 ± 0.058	0.858 ± 0.063	0.846 ± 0.053	0.842 ± 0.056	0.815 ± 0.037	0.829 ± 0.060	0.866 ± 0.032
slot_f1	0.483 ± 0.136	0.427 ± 0.088	0.362 ± 0.146	0.418 ± 0.124	0.283 ± 0.142	0.420 ± 0.142	0.390 ± 0.142
sementic_frame_acc	0.073 ± 0.081	0.044 ± 0.054	0.046 ± 0.052	0.051 ± 0.062	0.026 ± 0.070	0.070 ± 0.069	0.030 ± 0.083
loss	5.001 ± 1.760	5.950 ± 1.904	5.287 ± 1.946	6.140 ± 2.033	13.099 ± 4.899	5.417 ± 1.876	13.122 ± 5.131

Таблица 3: Таблица сравнения качества по языкам для атаки Word level

	en	de	es	fr	ja	pt	zh
intent_acc	0.919 ± 0.021	0.913 ± 0.024	0.913 ± 0.025	0.921 ± 0.023	0.897 ± 0.020	0.904 ± 0.029	0.890 ± 0.022
slot_f1	0.680 ± 0.142	0.661 ± 0.143	0.598 ± 0.147	0.636 ± 0.136	0.563 ± 0.068	0.658 ± 0.143	0.630 ± 0.051
sementic_frame_acc	0.305 ± 0.137	0.302 ± 0.116	0.192 ± 0.092	0.271 ± 0.117	0.037 ± 0.058	0.322 ± 0.151	0.054 ± 0.079
loss	2.949 ± 1.592	3.291 ± 0.956	3.628 ± 1.548	2.729 ± 0.828	10.237 ± 1.783	2.683 ± 1.314	9.786 ± 2.084

Таблица 4: Таблица сравнения качества по языкам для атаки Alignments

4 Заключение

AAAAAAAAAAAAAAAAAA FUCK ME

Список литературы

[1] Zi-Yi Dou and Graham Neubig. Word alignment by fine-tuning embeddings on parallel corpora. In *EACL*, 2021.