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:
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

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

3.1 Обучение моделей на датасете ATIS seven languages

В своей работе мы обучаем языковые модели решать задачу одновременной классификации интентов и разметки слотов в предложении. Эта задача заключается в определении желаемой цели запроса пользователя по предложению и классификации слов в предложении.

3.1.1 Датасет

В качестве датасета в своей работе мы выбрали датасет ATIS seven languages [6]. 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.

Каждый объект в датасете состоит из предложения, меток слов и интента. 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.

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 гипотезу о наличии кросс-язычных знаний у моделей. Тестовая выборка, которая будет нас интересовать в данном контексте состоит из всех семи языков, но мы оцениваем качество на каждом языке отдельно.

Каждая из моделей обучалась с одинаковыми гиперпараметрами - 10 эпох на обучающей выборке с длиной шага обучения 10^{-5} и размером батча в 64 объекта.

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 сме-

шения 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 [5]. 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 (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 Адверсариальная атака, общая схема

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. Этот вариант является грубой оценкой снизу, так как он не учитывает

контекста предложений и не учитывает многозначность слов.

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

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**
      token = $\mathbb{T}[\mathbb{L}]$[x[token_id]]
      x[token_id] = token
    end if
  return x, y
  end function

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 языков. Кандидаты для каждого токена определяются как токены из предложения на встраиваемом языке, в которые был выровнен токен.

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

Algorithm 3 Word-level атака

Require: Выравнивание предложения на исходном языке к предложению на целевом языке $\mathbb{A}$
  function GetCandidates($\mathcal{M}$, x, y, token_id)
    if x[token_id] in $\mathbb{A}[\mathbb{L}]$ **then**
      tokens = $\mathbb{A}[\mathbb{L}]$[x[token_id]]
      x[token_id] = tokens
      y[token_id] = ExtendSlotLabels(y[token_id], len(tokens))
    end if
  return x, y
  end function

3.3 Метод 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.

3.3.1 Метод адверсариального предобучения

3.4 Адверсариальные атаки на защищенные модели

В данной секции мы сравним результаты

3.4.1 Сравнение на тестовой выборке

	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

3.4.2 Сравнение для word-level атаки

	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 word-level атаки

	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 word-level атаки

3.4.3 Сравнение для phrase-level атаки

	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 phrase-level атаки

	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 phrase-level атаки

3.5 Результаты

3.5.1 Кросс-язычные знания в моделях

3.5.2 Качество моделей после адверсариальных атак

3.5.3 Влияние метода адверсариального предобучения

4 Заключение

AAAAAAAAAAAAAAAAAA FUCK ME

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)*. Minneapolis, Minnesota: Association for Computational Linguistics, июнь 2019, с. 4171—4186. DOI: 10.18653/v1/N19-1423. URL: https://www.aclweb.org/anthology/N19-1423.

[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] Shana Poplack, DAVID SANKOFF и CHRISTOPHER MILLER. «The social correlates and linguistic processes of lexical borrowing and assimilation». В: *Linguistics* 26 (янв. 1988), с. 47—104. DOI: 10.1515/ling.1988.26.1.47.

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