Intro to DeepPavlov. Multi-Task Learning.

Dmitry Karpov

Moscow Institute of Physics and Technology

Библиотека DeepPavlov

Как использовать библиотеку

Многозадачное обучение

Многозадачное обучение - исследования переноса знаний на диалоговых наборах данных

Библиотека DeepPavlov

Определение

Открытая библиотека обработки естественного языка DeepPavlov[1] предназначена для использования многоязычных и языко-специфичных NLP моделей как самостоятельно, так и в рамках диалоговых систем.

[1] Mikhail Burtsev и др. «DeepPavlov: An Open Source Library for Conversational AI». В: *NIPS*. 2018. URL: https://openreview.net/pdf?id=BJzyCF6Vn7.

Особенности библиотеки

- DeepPavlov основан на `json` конфигурационных файлах моделей
- DeepPavlov использует `transformers` от HuggingFace
- DeepPavlov поддерживает REST API и может использоваться в качестве сервиса
- DeepPavlov распространяется под лицензией Apache 2.0

Ссылки на DeepPavlov

- DeepPavlov Demo[Click]
- DeepPavlov GitHub[6k Stars]
- DeepPavlov Forum
- DeepPavlov Documentation

Поддерживаемые модели

- Named-Entity Recognition
- Classification (sentiment / emotions / etc)
- Question Answering
- GLUE/SuperGLUE
- Russian SuperGLUE
- Few-shot models
- Multi-Task models
- Sentence ranking / sentence segmentation / etc

Как использовать библиотеку

Пример

Если кому интересно: вот конфигурационные файлы для поддерживаемых моделей

Russian Paraphraser Russian NER English NER Russian SQUAD English SQUAD Russian Sentiment English Sentiment Russian relation ranking Russian ODQA English ODQA Russian KBQA English KBQA

Многозадачное обучение - зачем оно нужно

- Экономия вычислительных ресурсов
- Рост качества на схожих задачах

Многозадачное обучение - как оно реализовано (энкодер-агностичные модели)

В 2 словах - отдельный задаче-специфичный линейный слой на каждую задачу, который применяется к выходу энкодера.
Достоинства архитектуры:

- Вычислительная и архитектурная простота
- Расширяемость на различные типы задач
- Не требует псевдоразметки
- Можно быстро заменить энкодер

Архитектура интегрирована в open-source библиотеку DeepPavlov. Пример

Многозадачное обучение - результаты на GLUE

htbp

Таблица: Метрики многозадачной энкодер-агностичной модели для набора задач GLUE. M.Corr означает корреляцию Мэттью, P/S означает корреляцию Пирсона-Спирмена, Acc точность, F1 - макро-F1. Режим S означает однозадачные модели, режим M означает многозадачные модели. Размер означает размер тренировочного набора данных.

Модель	Режим	Среднее	CoLA	SST-2	MRPC	STS-B	QQP	MNLI	QNLI	RTE	AX
		Размер	8.6k	67.3k	2.5k	5.7k	363.8k	392.7k	104.7k	2.5k	как у MNLI
		метрика	M.Corr	Acc	F1/Acc	P/S Corr	F1/Acc	Acc (m/mm)	Acc	Acc	M.Corr
Человек	-	87.1	66.4	97.8	86.3/80.8	92.7/92.6	59.5/80.4	92.0/92.8	91.2	93.6	-
distilbert	S	73.3	**42.4**	**92.1**	85.6/**80.3**	78.8/76.8	**69.5/88.5**	**81.3/80.8**	**87.5**	52.1	29.9
	M	**74.5**	36.0	91.0	**85.7**/79.9	**82.6/81.6**	68.4/87.4	80.4/80.3	86.0	**69.5**	**30.1**
bert	S	77.3	**53.7**	**93.2**	**87.7/82.8**	83.8/82.2	**70.3/88.9**	**83.8/83.1**	**90.6**	62.1	32.1
	M	**77.8**	45.8	92.9	86.8/82.2	**85.3/84.7**	70.2/88.6	83.5/82.6	90.1	**74.5**	**32.8**
bert-large	S	**79.5**	**59.2**	**94.9**	85.0/80.6	**85.8/84.5**	70.5/89.1	**86.7/85.6**	92.2	70.1	**39.4**
	M	**79.5**	50.8	94.1	**87.3/82.8**	83.8/83.9	**71.0/89.2**	85.9/85.0	**92.4**	**78.5**	38.5

Энкодер-агностичные модели: данные, сэмплирование

Принцип подбора данных:

- Разговорные задачи
- Совпадающие классы для английского и русского языка

Сэмплирование примеров на каждом этапе обучения - батч из каждого набора данных с вероятностью пропорционально размеру (plain sampling).

Энкодер-агностичные модели: данные

- Для классификации **эмоций** – русскоязычный набор данных CEDR, собранный из различных интернет-источников, и англоязычный набор данных go_emotions, собранный из комментариев на ресурсе «Реддит». Использовалось семь типов эмоций по Экману – ярость, страх, грусть, удовольствие, удивление, отвращение, нейтральная.
- Для классификации **тональности** – англоязычный набор данных DynaSent(r1), состоящий из предложений, возникающих в диалогах, и русскоязычный набор данных RuReviews, состоящий из отзывов крупного российского электронного магазина. Использовалось три класса – положительный, отрицательный, нейтральный.

Энкодер-агностичные модели: данные

- Для классификации **токсичности** – русскоязычный набор комментариев с ресурса «Двач» (RuToxic) и англоязычный набор комментариев из Википедии (Wiki Talk). Использовалось два класса – токсичный и не токсичный.
- Для классификации **тем и классификации интентов** – набор данных MASSIVE, состоящий из обращенных к диалоговой системе фраз пользователей. Набор существует и использовался как в англоязычном, так и в русскоязычном варианте. Каждая фраза из набора принадлежит к одной из 60 тем и к одному из 18 интентов.

Энкодер-агностичные модели: сравнение с однозадачными, английский язык

Таблица: Метрики англоязычных моделей (точность/макро-F1) для пяти англоязычных диалоговых задач.Режим S означает однозадачные модели, режим M означает многозадачные модели. Усреднено по трем запускам.

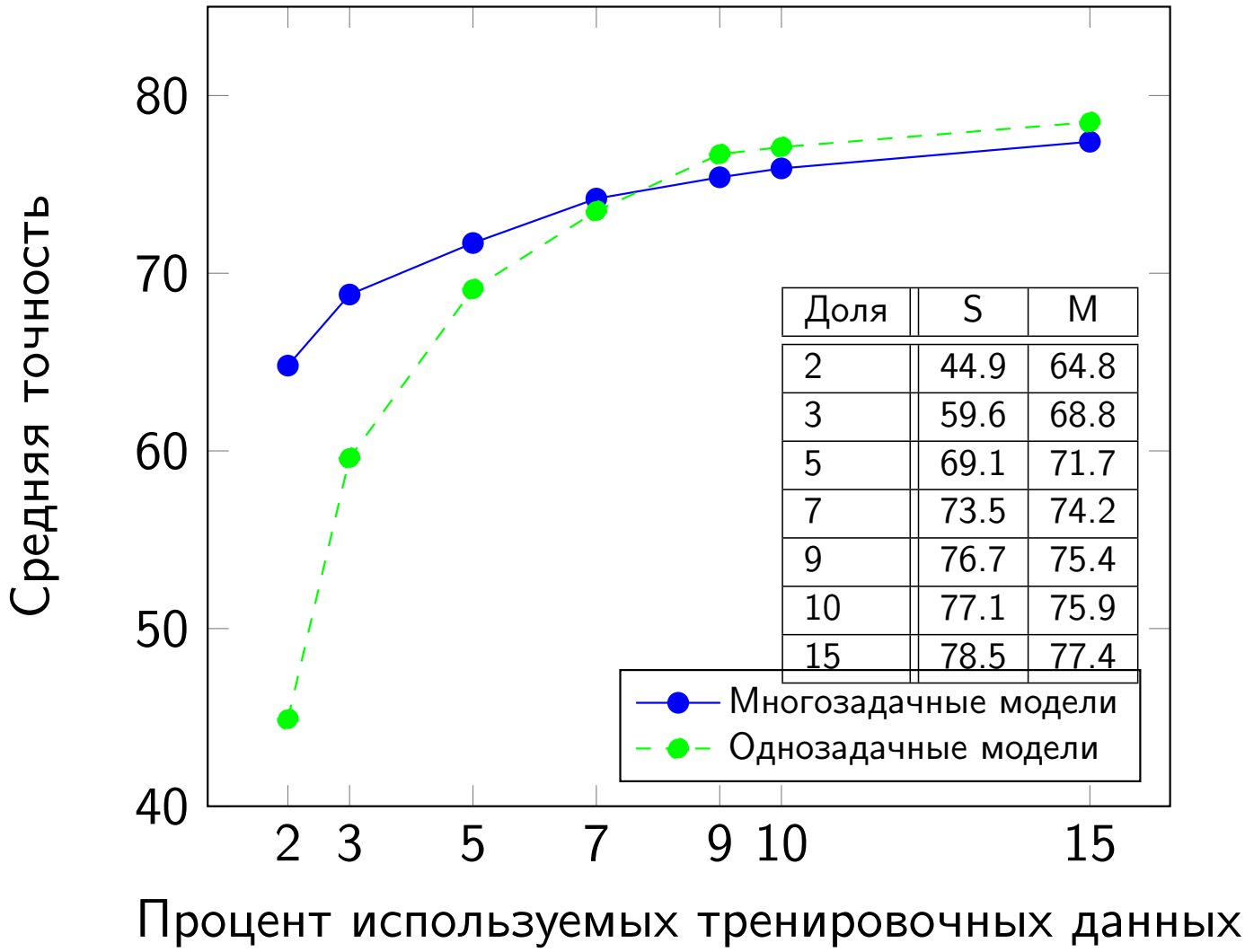
Модель	**Режим**	**Среднее**	Эмоции 39.4k	Тональность 80.5k	Токсичность 127.6k	Интенты 11.5k	Темы 11.5k
distilbert	S	**82.9**	**70.3**	74.7	91.5	**87.4**	**91.0**
	M	82.1	67.7	**75.2**	90.6	86.3	90.8
bert	S	**83.9**	**71.2**	76.1	**93.2**	**87.9**	**91.3**
	M	83.0	69.0	**76.5**	91.4	87.1	91.2
bert-large	S	**84.7**	**70.9**	**80.5**	**92.1**	**88.4**	91.3
	M	83.6	69.0	79.0	91.3	87.3	**91.3**

Энкодер-агностичные модели: сравнение с однозадачными, русский язык

Таблица: Метрики русскоязычных моделей (точность/f1 macro) для пяти диалоговых задач. Режим S означает однозадачные модели, режим M означает многозадачные модели. Усреднено по трем запускам.

Модель	**Режим**	**Среднее**	Эмоции 6.5k	Тональность 82.6k	Токсичность 93.3k	Интенты 11.5k	Темы 11.5k
distilrubert	S	86.9	82.2	77.9	97.1	86.7	90.4
	M	86.3	81.0	77.7	96.9	85.2	90.7
rubert	S	86.5	80.9	78.0	97.2	86.2	90.0
	M	86.2	80.5	77.6	96.8	85.3	90.5

Энкодер-агностичные модели - эффект уменьшения размера выборки, данные на английском языке



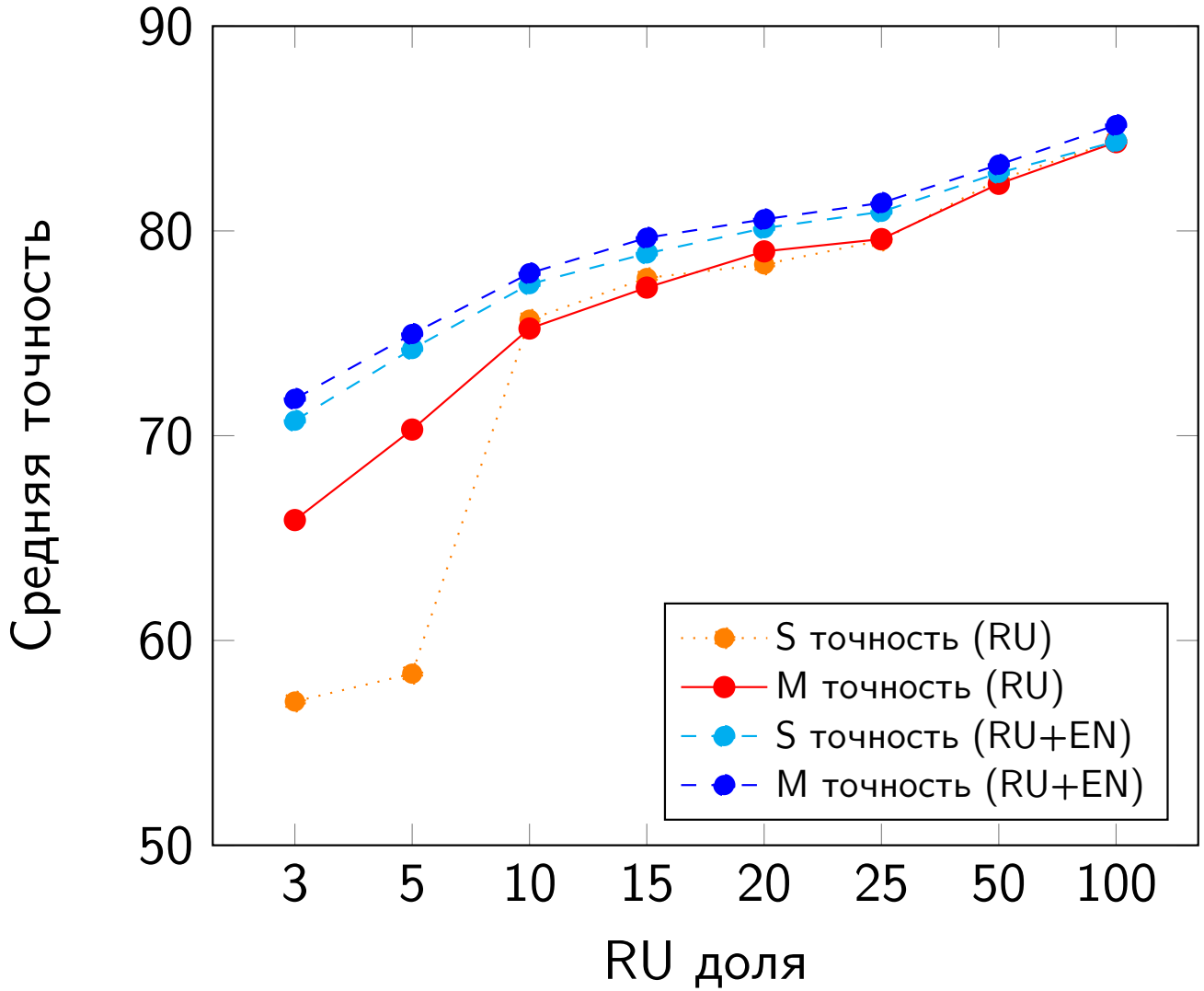
Доля	S	M
2	44.9	64.8
3	59.6	68.8
5	69.1	71.7
7	73.5	74.2
9	76.7	75.4
10	77.1	75.9
15	78.5	77.4

Энкодер-агностичные модели - эффект уменьшения размера выборки, данные на русском языке

Таблица: Средняя точность многозадачных моделей в зависимости от того, на какой доле русскоязычных данных они обучались, и того, добавлялись ли к ним англоязычные данные, для многоязычного distilbert.

RU доля	S RU	M RU	S RU+EN	M RU+EN
3	57.0	**65.9**	**71.8**	70.7
5	58.4	**70.3**	**75.0**	74.2
10	**75.7**	75.2	**77.9**	77.4
15	**77.7**	77.2	**79.7**	78.9
20	78.4	**79.0**	**80.6**	80.1
25	79.5	**79.6**	**81.4**	80.9
50	**82.5**	82.3	**83.2**	82.8
100	**84.4**	84.3	**85.2**	84.4

Энкодер-агностичные модели - эффект уменьшения размера выборки, данные на русском языке



Энкодер-агностичные модели: выводы

1. Многозадачные энкодер-агностичные модели - почти как однозадачные. Если для какой-то задачи данных мало, но для другой похожей задачи их много - то даже лучше.
2. Если данных становится очень мало, то многозадачные модели становятся сильно лучше однозадачных. Опять же, зависит от размера данных для задачи.
3. Добавление английских данных к русским - улучшает метрики, чем меньше русских данных - тем сильнее (до нескольких процентов). Это верно и для однозадачных моделей, при любом языке валидации.

Спасибо за внимание