tasksource: A Large Collection of NLP tasks with a Structured Dataset Harmonization Framework

Damien Sileo damien.sileo@inria.fr

LREC-COLING 2024

https://github.com/sileod/tasksource

https://huggingface.co/tasksource

Outline

Context

Dataset preprocessing

Contributions

A preprocessing / dataset parsing framework
Dataset parsing annotations
Pre-trained models
Recasting tools and dataset (tasksource instruct)

Motivation

- Combinining datasets works well (MTL, instruction-tuning)

Motivation

- Combinining datasets works well (MTL, instruction-tuning)

- Combining structures and column names are a hurdle

hf.co quora

questions sequence	is_duplicate bool
{ "id": [1, 2], "text": ["What is the step by step guide to invest in share market in india?", "What is the step by step guide to invest in share market?"] }	false

hf.co glue/mrpc

sentence1 string · *lengths*	sentence2 string · *lengths*	label class label
In Paris , in October 1560 , he secretly met the English…	In October 1560 , he secretly met with the English ambassador ,…	0 0

Motivation

- Combinining datasets works well (MTL, instruction-tuning)

- Combining structures and column names are a hurdle

→ Meet tasksource

hf.co quora

questions sequence	**is_duplicate** bool
{ "id": [1, 2], "text": ["What is the step by step guide to invest in share market in india?", "What is the step by step guide to invest in share market?"] }	false

hf.co glue/mrpc

sentence1 string · *lengths*	**sentence2** string · *lengths*	**label** class label
In Paris , in October 1560 , he secretly met the English…	In October 1560 , he secretly met with the English ambassador ,…	0 0

Anatomy of a dataset preprocessing

We define *Dataset parsing* is mapping a dataset to a template

Considered templates:

- **Classification**(text1, text2, label)
- **MultipleChoice**(prompt, choices, label)
- **TokenClassification**(tokens, labels)
- **Seq2Seq(**text1,text2**)**

functions

Anatomy of a dataset preprocessing

Classification(text1, text2, label)

questions sequence	is_duplicate bool
{ "id": [1, 2], "text": ["What is the step by step guide to invest in share market in india?", "What is the step by step guide to invest in share market?"] }	false
{ "id": [3, 4], "text": ["What is the story of Kohinoor (Koh-i-Noor) Diamond?", "What would happen if the Indian government stole the Kohinoor (Koh-i-Noor) diamond…	false

fields

```
quora = Classification( Template
  text1=lambda x: x:x['questions'][0], functions
  text2=lambda x: x:x['questions'][1],
  label=lambda x:x['is_duplicate']
)
```

Anatomy of a dataset preprocessing

Classification(text1, text2, label)

```
quora = Classification(
  text1=get.questions.text[0],
  text2=get.questions.text[1],
  labels='is_duplicate')
```

tasksource syntax (shorter)

“get” utility

values treated as constant functions

```
quora = Classification(
  text1=lambda x: x:x['questions'][0],
  text2=lambda x: x:x['questions'][1],
  label=lambda x:x['is_duplicate']
)
```

Template fields

functions

Usage

```
pip install tasksource
```

```
quora = tasksource.Classification(
    get.questions.text[0],
    get.questions.text[1],
    'is_duplicate'
    dataset_name="quora" # HuggingFace 🤗 path
)

dataset = quora.load()
```

Usage

```
quora = tasksource.Classification(
    get.questions.text[0],
    get.questions.text[1],
    'is_duplicate'
    dataset_name="quora" # HuggingFace 🤗 path
)

dataset = quora.load()

# quora and hundreds of tasks are already coded
dataset = tasksource.load_task('glue/qqp')
```

Usage

```
quora = tasksource.Classification(
    get.questions.text[0],
    get.questions.text[1],
    'is_duplicate'
    dataset_name="quora" # HuggingFace 🤗 path
)

dataset = quora.load()

# quora and hundreds of tasks are already coded
dataset = tasksource.load_task('glue/qqp')
```

using tasksource:
-concise
-transparent
-reproducible

Tasksource tasks (provided preprocessings)

English subset: 600 tasks

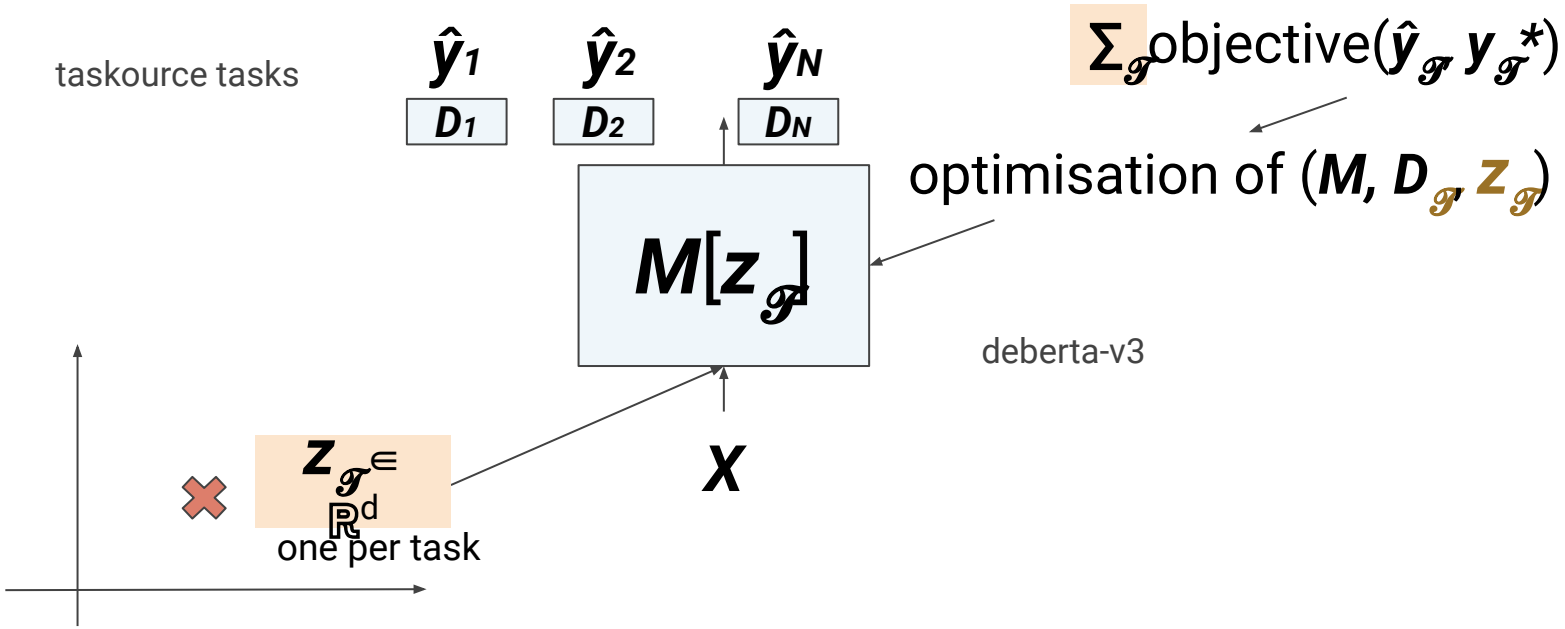
including many NLI tasks, reasoning tasks, discourse understanding multiple choice, commonsense, emotion analysis…

Multilingual subset: 400 tasks

I only picked multilingual tasks, i.e. no monolingual task

Still many datasets to annotate

Multi-task tuning on tasksource



[Achille et al. 2019, Pilault et al. 2019, Sileo et al. 2022]

Multi-task tuning on tasksource

B. Model Recycling results

model_name	deberta-v3-base	+tasksource
avg	79.04	**80.73**
mnli (linear probe)	-	**93.73**
20_newsgroup	86.41	**86.46**
ag_news	90.44	**90.67**
amazon_reviews_multi	66.86	**66.90**
anli	58.78	**60.38**
boolq	82.99	**85.66**
cb	75.00	**82.14**
cola	86.57	**87.15**
copa	58.40	**81.00**

Available on HuggingFace as a backbone or NLI / zero-shot classifier: hf.co/sileod/deberta-v3-base-tasksource-nli

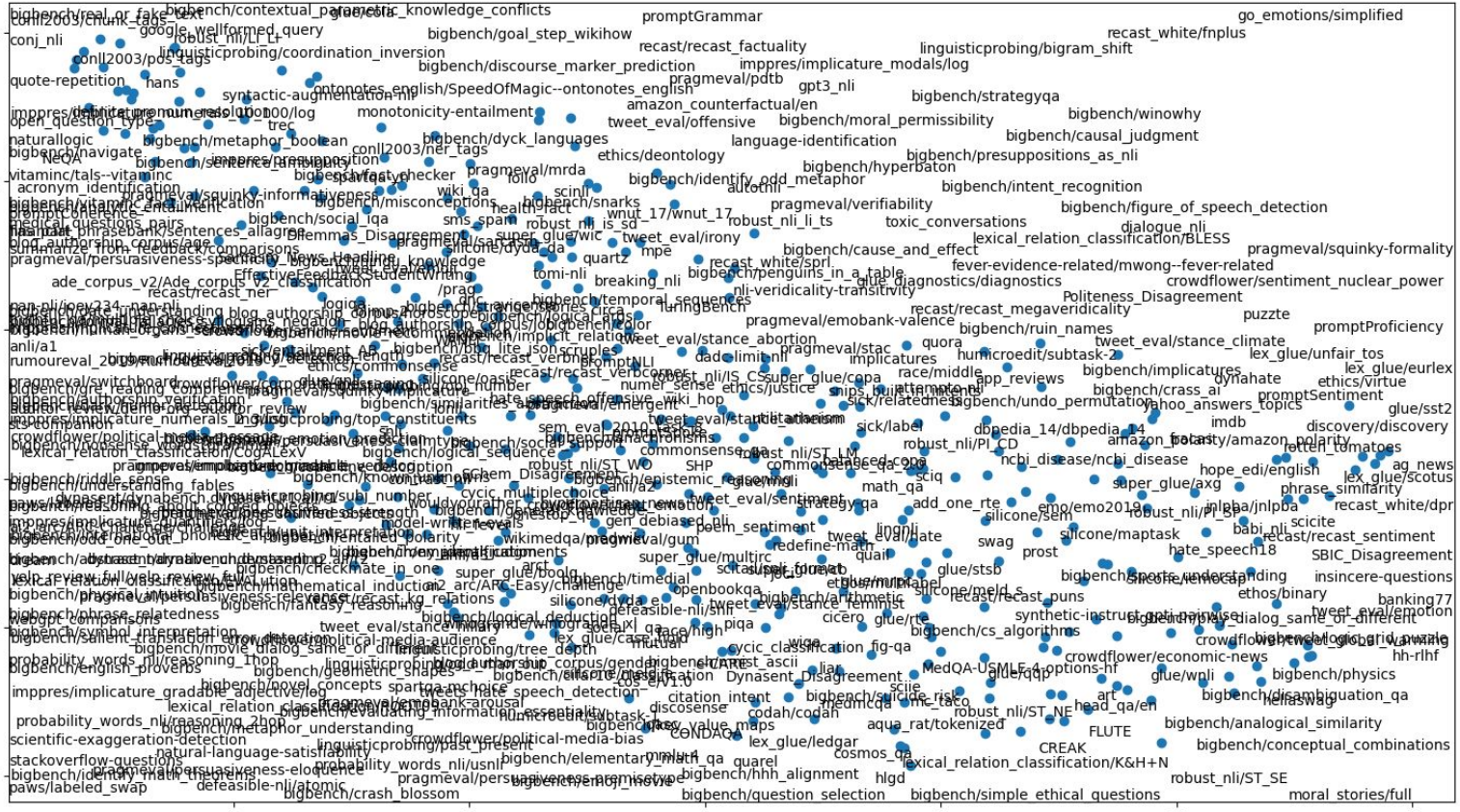
Ranked **1st** 🥇 across **2685** models on IBM NLU benchmark https://ibm.github.io/model-recycling/

Also available in large/small/multilingual

and paraphrase/sentiment/toxicity/RLHF specialized models

https://github.com/sileod/tasknet (training code)

Task embeddings



Tasksource recast

🤗/zero-shot-label-nli

labels	premise	hypothesis	task
2 contradiction	Only a Liberal Rag would publish such drivel Winning	This example is not hate-speech.	mbib-base/hate-speech

🤗/tasksource-instruct-v0

inputs	targets	task
With no explanation, label A→B with either "entailment", "neutral" or "contradiction". A: After learning that one of its members had been taken in by the scheme, the Middle East Studies Association posted a warning on its Web site. B: A member of the Middle East Studies Association was scammed for money.	neutral.	glue/mnli

https://github.com/sileod/tasksource
https://github.com/Data-Provenance-Initiative

https://huggingface.co/tasksource

Future work:
-automating dataset parsing
-keep adding more tasks (open to contribution)