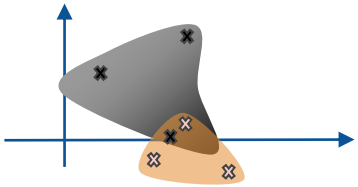
KU LEUVEN

Analysis and prediction of NLP models via task embeddings



Damien Sileo and Marie-Francine Moens
KU Leuven
damien.sileo@kuleuven.be

Calculus

Horizon 2020
Advanced Grant 788506

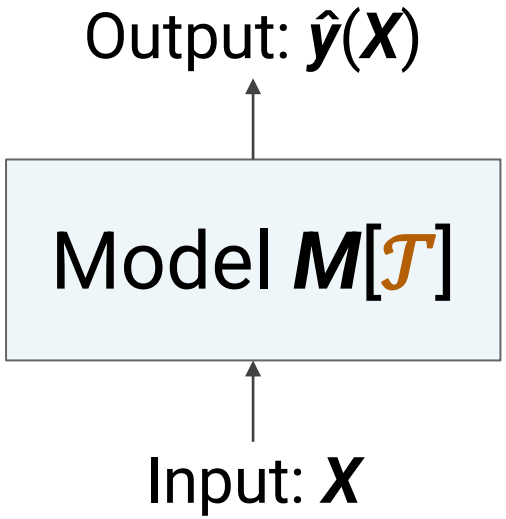
Introduction

Transfer & task embeddings

Contributions

New ressource: **metaEval**
Experimental setting
Analyses of task embeddings

Task



Task $\mathcal{T}$

$\{ (X, y^*) \in \mathcal{D}_{\mathcal{T}} \}$ — *Dataset*

$\text{objective}_{\mathcal{T}}: \hat{y}(X), y^* \mapsto \mathbb{R}$ — *Objective (e.g. log-likelihood)*

Transfer between tasks

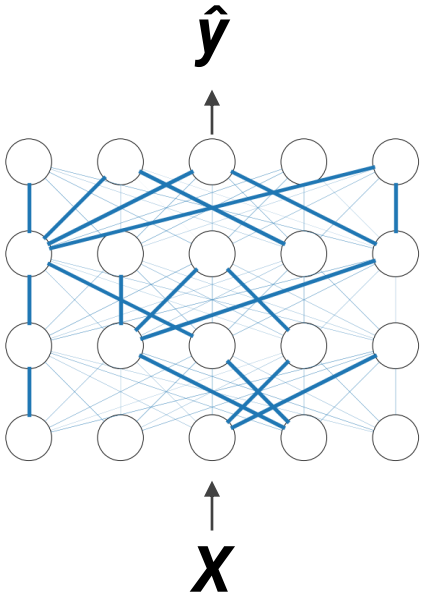
M[∅ → RTE]

<

M[∅ → MNLI → RTE]

SOURCE TARGET

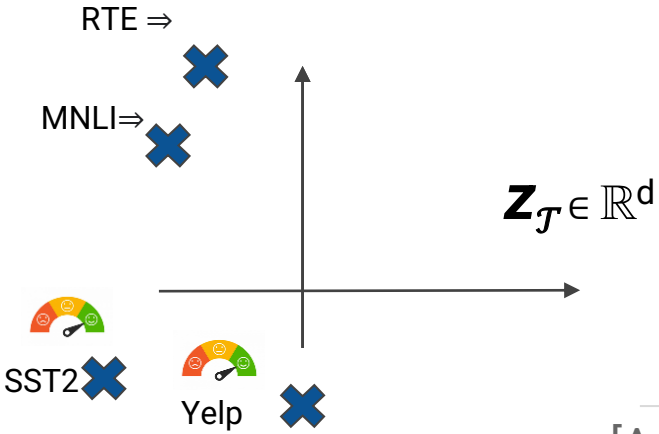
transfer



objective(ŷ, y*)

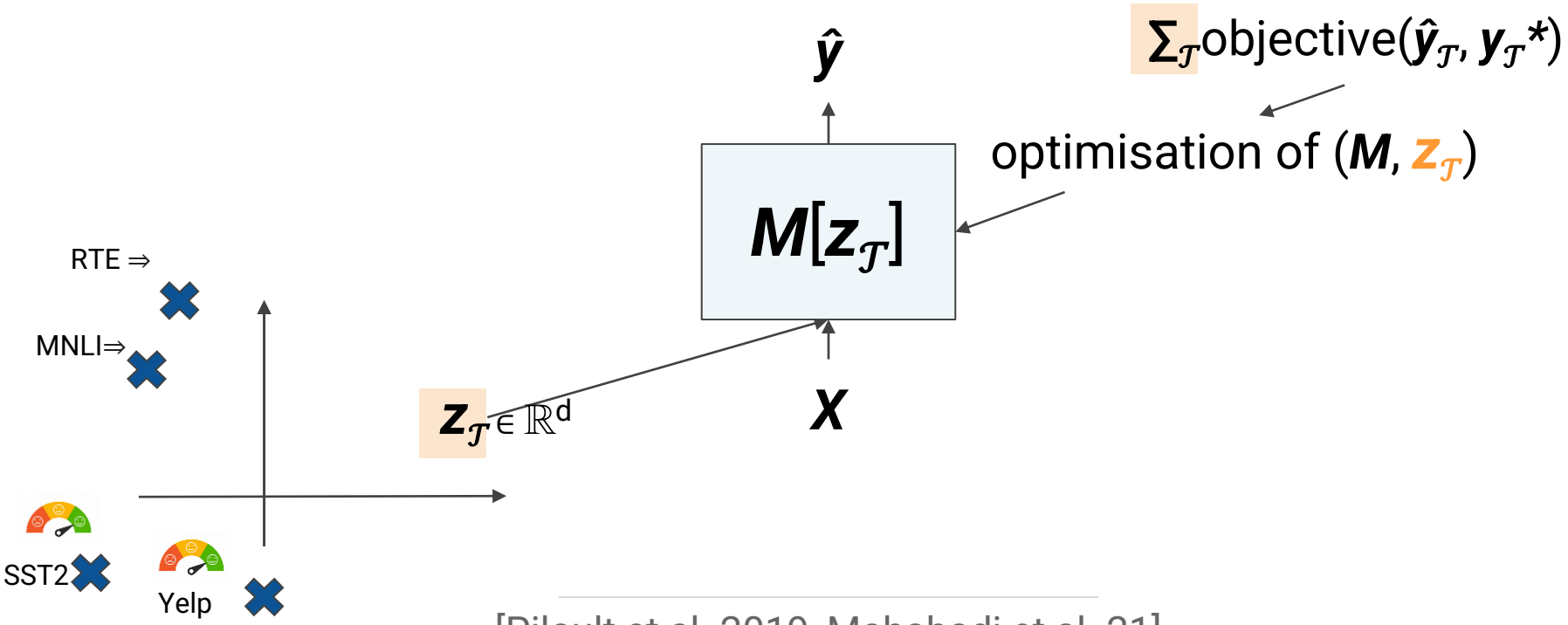
optimisation

Task embeddings



[Achille et al. 2019, Vu et al. 2019]

Learned task embeddings



[Pilault et al. 2019, Mahabadi et al. 21]

Introduction

Transfer & task embeddings

Contributions

New resource: **metaEval**

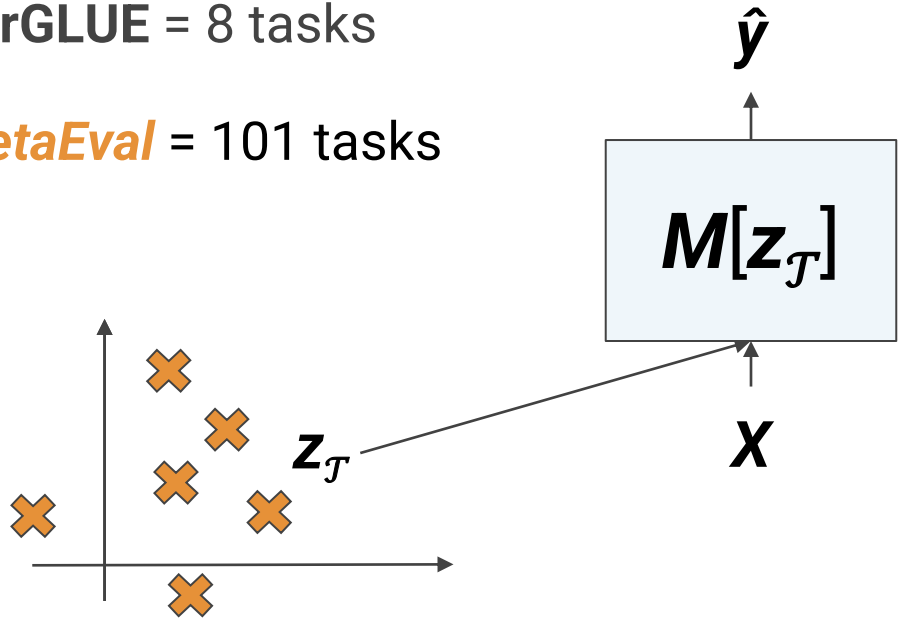
Experimental setting

Analyses of task embeddings

Collections of tasks

GLUE = 9 tasks
SuperGLUE = 8 tasks

MetaEval = 101 tasks



MetaEval

Huggingface Datasets (Wolf et al. 20)

GLUE (Wang et al. 19)

SuperGLUE (Wang et al. 19)

Other classification tasks

\+ pragmEval (Sileo et al. 21)

\+ TweetEval (Barbieri et al. 20)

\+ Recast (Poliak et al. 18)

\+ Blimp (Warstadt et al. 20)

\+ Ethics (Hendrycks et al. 21)

\+ CrowdFlowers (Van pelt et al. 12)

101 tasks

Unified 🤗 dataset format: (train/val/test)

sentence1	sentence2	label
sentence	-	label

```
dataset = load_and_align('health_fact')
```

https://github.com/sileod/metaeval

Introduction

Transfer & task embeddings

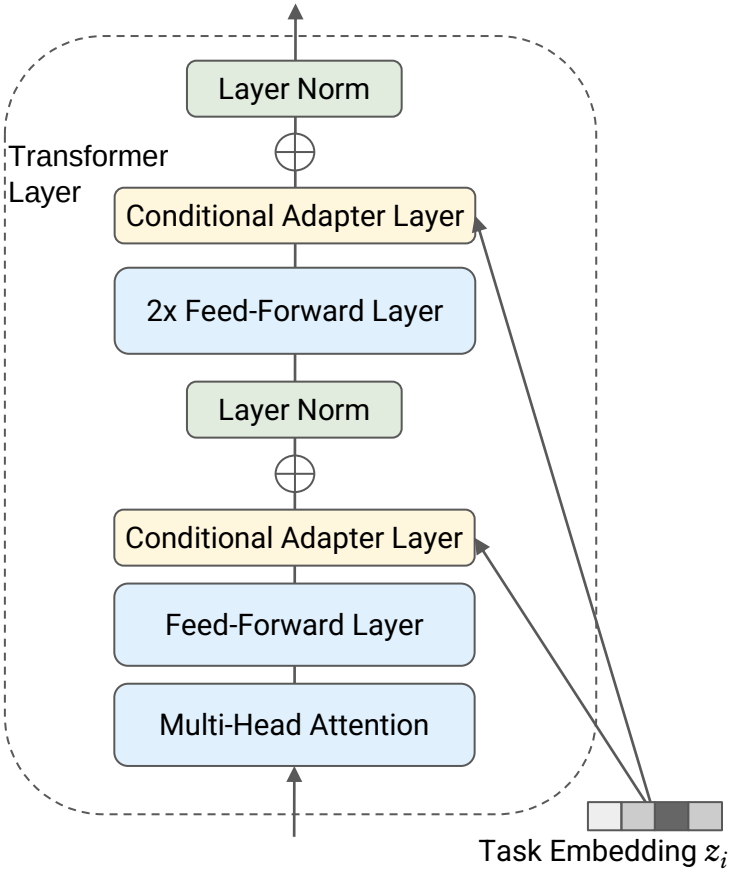
Contributions

New resource: **metaEval**

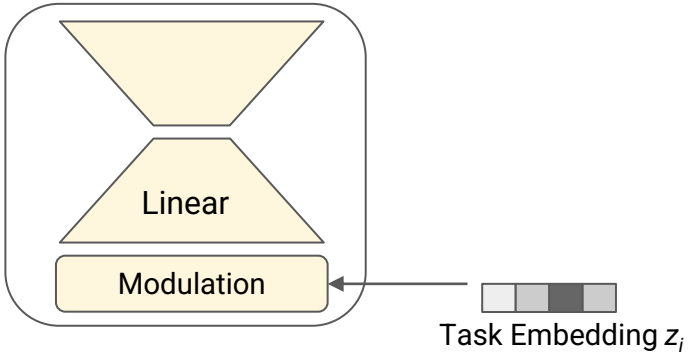
Experimental setting

Analyses of task embeddings

Conditional Adapters



Conditional Adapter Layer



[Pilault et al. 21]

trained

frozen

Evaluation

Fine-tuning method	MetaEval test accuracy	Trained encoder parameters	Task-specific trained encoder parameters
Majority Class	42.9	-	-
Full Fine-Tuning (1 model/task)	76.9	124M	124M
Adapter (1 adapter/task)	67.8	10M	10M
Conditional Adapter (Mahabadi et al. 2021)	75.6	38M	512
Conditional Adapter (Pilault et al .2021)	**79.7**	10M	32
z=TextEmb task embedding (Vu et al. 2020)	69.9	10M	
z=Fisher-information task embedding (Fisher et al 2020)	67.5	10M	

Introduction

Transfer & task embeddings

Contributions

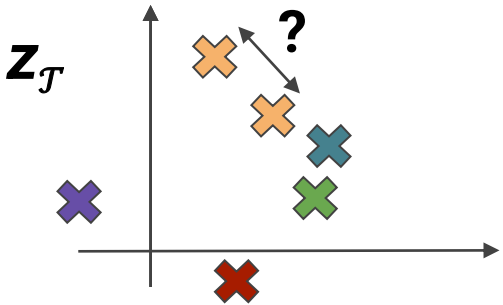
New resource: **metaEval**

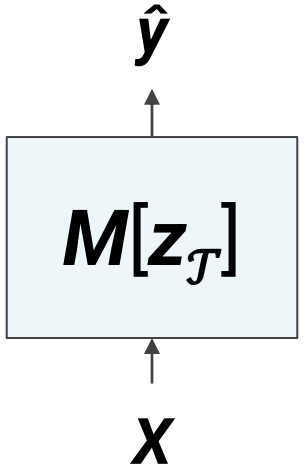
Experimental setting

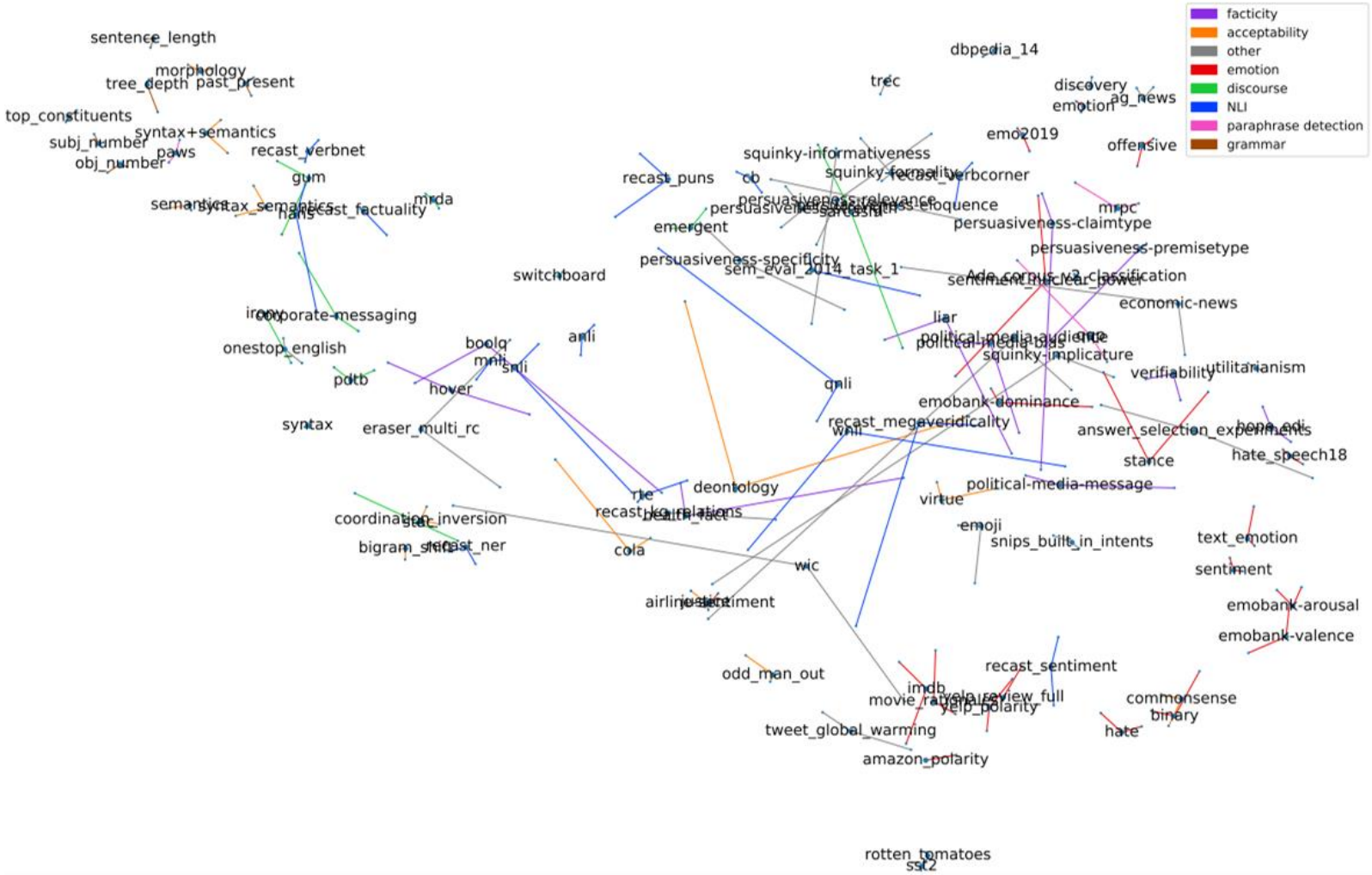
Analyses of task embeddings

Stability of task embeddings

duplicated task
2 representations for an identical task
= proximity ?

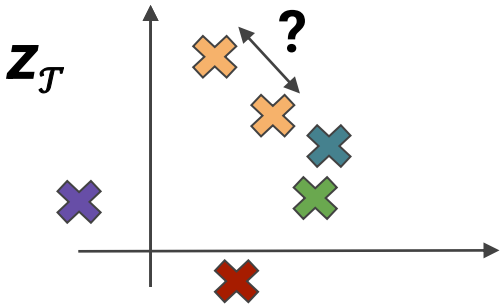


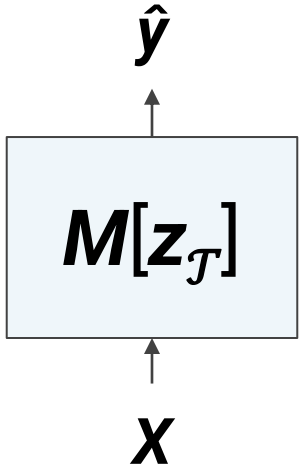



facticity
acceptability
other
emotion
discourse
NLI
paraphrase detection
grammar
sentence_length
morphology
tree_depth
past_present
top_constituents
syntax+semantics
subj_number
paws
recast_verbnet
obj_number
gum
semantics
syntax_semantics
mrda
hans
recast_factuality
irony
corporate-messaging
onestop_english
boolq
mnli
snli
anli
pdtb
hover
syntax
eraser_multi_rc
switchboard
dbpedia_14
trec
discovery
ag_news
emotion
emo2019
offensive
squinky-informativeness
squinky-formality
recast_verbcorner
recast_puns
cb
persuasiveness-relevance
persuasiveness-eloquence
sarcasm
mrpc
emergent
persuasiveness-claimtype
persuasiveness-premisetype
persuasiveness-specificity
sem_eval_2014_task_1
Ade_corpus_v2_classification
sentiment_nuclear_power
economic-news
liar
political-media-audience
political-media-bias
squinky-implicature
verifiability
utilitarianism
cnli
emobank-dominance
recast_megaveridicality
wnli
answer_selection_experiments
hope_edi
hate_speech18
stance
political-media-message
virtue
rte
deontology
recast_kg_relations
health_fact
coordination_inversion
sst2
bigram_shift
recast_ner
cola
wic
emoji
snips_built_in_intents
text_emotion
sentiment
airline_sentiment
emobank-arousal
emobank-valence
odd_man_out
recast_sentiment
imdb
movie_rationales
yelp_review_full
yelp_polarity
commonsense
binary
hate
tweet_global_warming
amazon_polarity
rotten_tomatoes
ssc2

Stability of task embeddings

duplicated task
2 representations for an identical task
= proximity ?

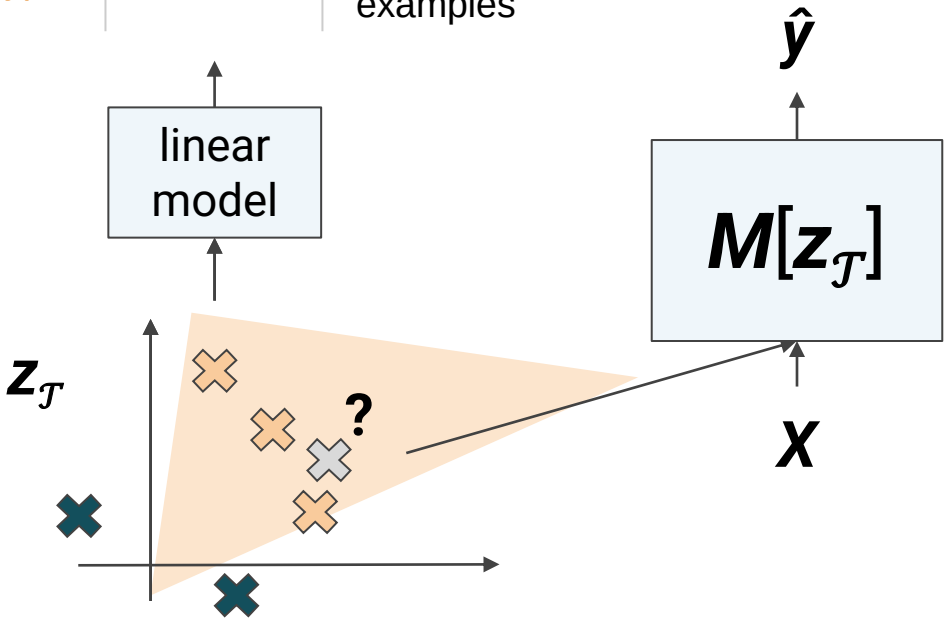




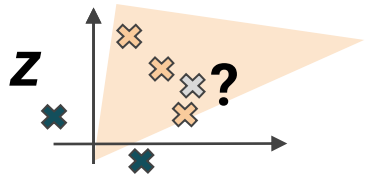
Task type	Position stability
Grammar	62.0 ± 3.9
Acceptability	57.1 ± 0.0
Emotion	47.6 ± 2.2
Discourse	45.7 ± 0.0
NLI	37.5 ± 1.0
Other	34.8 ± 0.7
Paraphrase detection	31.5 ± 13.1
Facticity	30.0 ± 4.7
Random embedding	1.0 ± 0.5

Probing of task embeddings

task type | domain | number of examples



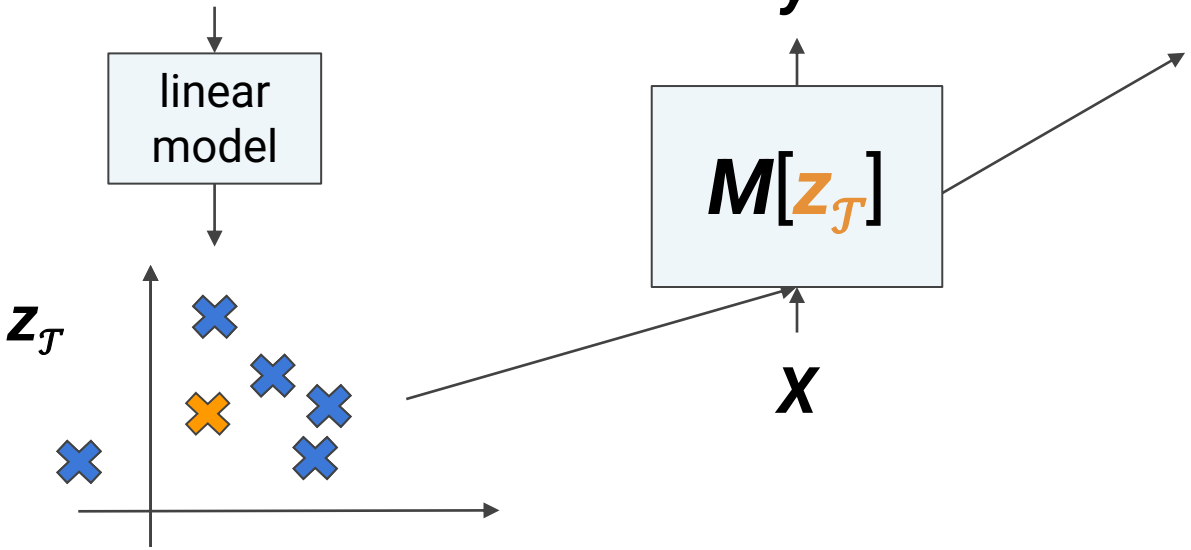
Probing task embeddings



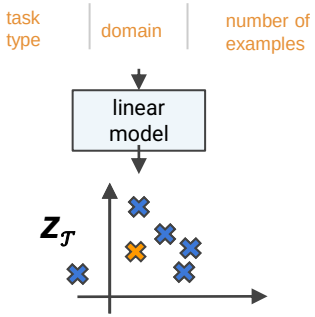
Feature	Domain-Cluster	Num-Rows	Num-Text-Fields	Task-Type	Text-Length
n.a. (Majority Class)	27.8	**62.4**	63.4	20.8	24.8
Fisher Embedding	37.7	61.3	62.2	35.7	25.6
Learned Features	41.8	51.5	62.2	45.7	32.8
Learned Features⊕TextEmb	71.5	60.4	76.3	**60.4**	53.6
TextEmb	**78.3**	59.3	**82.3**	50.6	**60.3**

Prediction of task embeddings

task type | domain | number of examples



Prediction of task embeddings



Feature	CoLa	SST2	MRPC	QQP	MNLI	QNLI	RTE	AVG
Single-Task Full-Fine-Tuning (Supervised)	79.2	93.1	75.5	84.7	80.9	88.9	47.3	78.5
Same Task-Type Full Fine-Tuning	73.5	**93.6**	68.8	55.3	**72.7**	51.5	**70.2**	69.4
Features-Aware Task Embeddings - Aspects (ours)	75.4	90.0	**70.4**	**71.1**	66.2	56.2	63.7	**70.4**

Conclusion

MetaEval : 101 aligned text classification datasets

Task embeddings lie in **regions** of space

Task embedding can predict **aspects**

Aspects can predict **task embedding**

damien.sileo@kuleuven.be