University of Zurich[UZH]

Department of Sociology & Department of Computational Linguistics

Evaluation of Transfer Learning and Domain Adaptation for Analyzing German-Speaking Job Advertisements

Ann-Sophie Gnehm, Eva Bühlmann & Simon Clematide

LREC 2022, 20 – 25 June

Motivation

- Text mining for Social Science on German-speaking job ads from 1990 to today:

  → How have job tasks and skill requirements developed in Switzerland over the last 30 years?

- Job ads:
  - Particular text type regarding structure and formulation
  - Rapid data shift over time

- Experiment with:
  - Transfer learning approaches: large-scale pre-training of LMs, only limited fine-tuning necessary
  - Domain adaptation techniques

- Share domain-adapted LMs and dataset

3 Tasks: Classification, Text Zoning, ICT Term Recognition

We are looking for a service oriented and dynamic
Global Mobility Manager
for XY Inc. XY Inc. offers you
…..

Your responsibilities include:
- …
- ...

What we expect:
- Degree in business, economics or law
- Flair for IT systems, proficiency in excel
- Service orientation and experience in a fast paced, high volume administration environment

Please send your application to …

1. Text classification:

- Profession? (34 classes)
- Main task? (21 classes)
- Experience needed? (3 classes)

2. Text zoning

- Token-level sequence labeling
- into 8 different text zone classes

3. ICT term recognition

- Formalized as NER-style task
- Recognize special domain terms

Continued In-Domain Pretraining of Language Representation Models

- General-domain pre-trained models:
  - **BERT-de:** bert-base-german-cased[1], trained on 12 GB data
  - **GBERT:** gbert-base[2], trained on 160 GB data

- **Domain vocabulary insertion for BERT-de**:
  - build domain-specific vocabulary with SentencePiece[3]
  - fill 3k empty spots in BERT-de vocabulary with most frequent subtokens:, e.g.:

    '#diplom' ('diploma'), 'Muttersprache' ('first language'), 'SAP'

1 https://huggingface.com/bert-base-german-cased; 2 Chan et al., (2020); 3 Kudo & Richardson (2018)

Continued In-Domain Pretraining of Language Representation Models (2/2)

– **Sampling of domain corpora:**

data source	data type	time span	data total			training data per epoch		
			# ads	# chars	size	# ads	# chars	size
SJMM corpus	representative sample	1990-2021	54K	67M	65MB	80K	93M	92 MB
OA corpus	large-scale web scraped	2014-2021	2.2M	3.9B	3.5GB	85K	150M	140 MB

→ upsampling SJMM, downsampling OA corpus (balanced in each epoch)

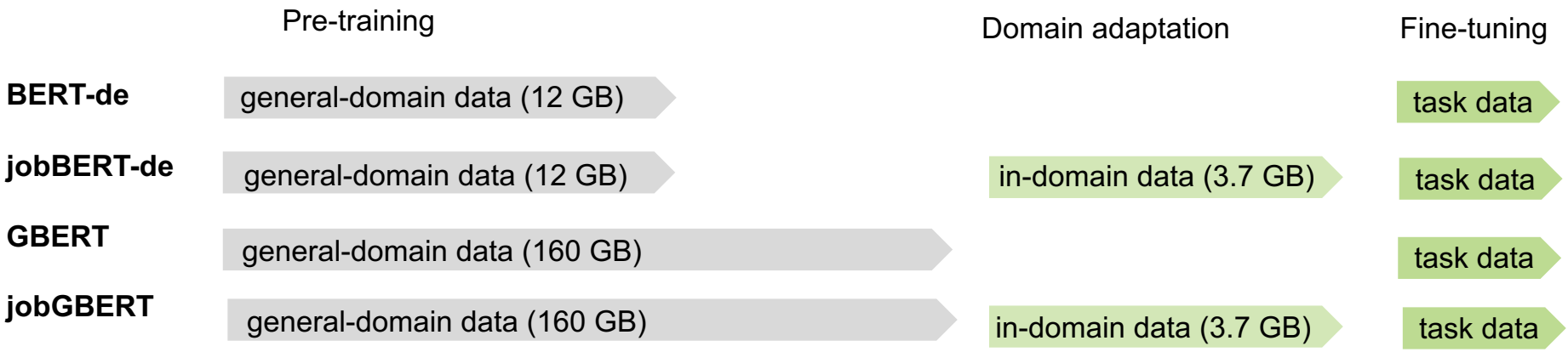
→ training over 25 epochs, exchanging OA material every epoch

– **Training parameters:**

- batch size 256, initial learning rate 5E-05, max. seq. length 512, for further in-domain training of BERT-de
- batch size 128, initial learning rate 1E-05 for further in-domain training of GBERT

Evaluation on Downstream NLP Tasks

– 4 language models:



– Fine-tuning on 3 downstream NLP tasks:
  – report performance estimates over 3-5 runs
  – Evaluation measures:
    – F1 for ICT term recognition task
    – imbalanced classes in classification and zoning task → balanced accuracy (Macro-Recall)[1]

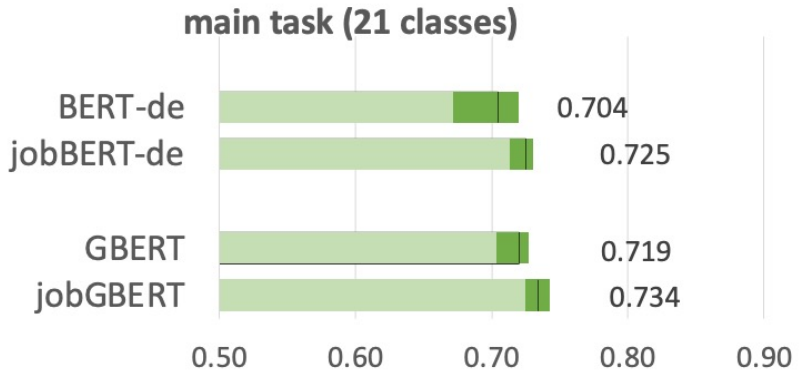
1 Grandini et al. (2020)

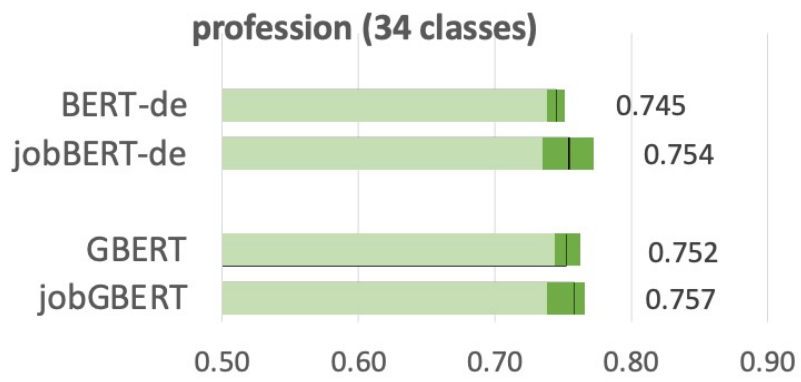
1. Classification Tasks

3 document classification tasks:

job ad text	... we are looking for an experienced pharmaceutical assistant as sales consultant ..
profession	32 - Medical, pharmaceutical professions
main task	7 - customer service, sales, cashier
experience	1 - needed

Balanced Accuracy for 3 classification task (n= 25k job ads, 80% train / 10% dev / 10% test set):





2. Text Zoning Task

Token-level sequence labeling task into 8 classes:

For/z1 our/z1 innovative/z1 product/z1 portfolio/z1 we/z3 are/z3 looking/z3 for/z3 an/z6 interior/z6 designer/z6 ./z6 You/z3 offer/z3 :/z3 -/z7 solid/z7 vocational/z7 training/z7 and/z7 experience/z7 ,/z7 -/z8 creativity/z8 and/z8 versatility/z8 We/z3 offer/z3 :/z3 -/z6 a/z6 high/z6 degree/z6 of/z6 autonomy/z6 ,z6 -/z6 a/z6 large/z6 studio/z6 ,/z6 -/z6 an/z6 interesting/z6 permanent/z6 position/z6 ./z6 ...

Zone	Definition	rel. Frequency
z1	company description	17.2%
z2	reason of vacancy	0.5%
z3	administration & residual text	25.2%
z4	job agency description	0.7%
z5	material incentives	1.7%
z6	job description	32.4%
z7	required hard skills	12.7%
z8	required personality (soft skills)	9.5%

Model	Balanced Accuracy		
	orig. train set 1990-2014 (n=22.5k), orig. test set 2010-2014 (n=650)	orig. train set 1990-2014 (n=22.5k), new test set 2015-2021 (n=150)	updated train set 1990-2021 (n=23.1k) , new test set 2015-2021 (n=150)
BERT-de	0.855	0.826	0.847
jobBERT-de	0.874	**0.894**	**0.941**
GBERT	0.869	0.859	0.885
jobGBERT	**0.880**	0.876	0.896

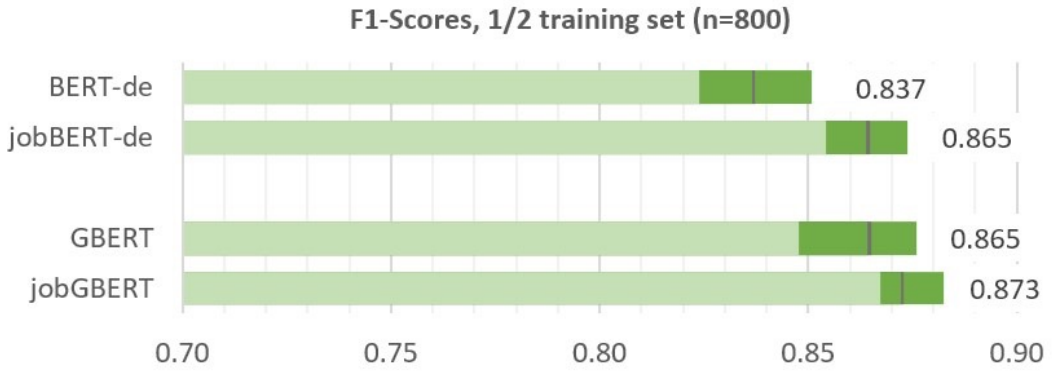
3. ICT Term Recognition

Formalized as NER-style task using spaCy:

> Several years of professional experience in a similar function, very good **PC skills** (**MS-Office** , especially **Word** and **Excel** , if possible, experience with **Abacus**) as well as stylistically confident written and spoken German are required.

Data Set:
2000 manually annotated job ads (German)

- 80% train / 10% dev / 10% test
- targeted sampling strategy for recall and coverage; based on a domain-specific topic model



Conclusions

- Domain adaption techniques:
  - Highly beneficial for all 3 tasks
  - Efficient: jobBERT-de competitive with GBERT
  - Stronger effects for models with less general-domain pre-training

- both more extensive general-domain pre-training and in-domain pre-training:
  - mitigate effect of data shift over time
  - help to deal with small training data sets

- Open questions:
  - Effect of domain-specific vocabulary extension
  - Hyper-parameter grid search or ensembling strategies for task-specific fine-tuning

References

Chan, B., Schweter, S., and Mo ̈ller, T. (2020). Ger- man's Next Language Model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirec- tional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computa- tional Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171– 4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Grandini, M., Bagli, E., and Visani, G. (2020). Met- rics for Multi-Class Classification: an Overview. *arXiv:2008.05756 [cs, stat]*, August. arXiv: 2008.05756.

Kudo, T. and Richardson, J. (2018). SentencePiece: A simple and language independent subword tok- enizer and detokenizer for Neural Text Processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.