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
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](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:**

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

→ 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:

<table>
<thead>
<tr>
<th>Pre-training</th>
<th>Domain adaptation</th>
<th>Fine-tuning</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT-de</td>
<td>general-domain data (12 GB)</td>
<td>task data</td>
</tr>
<tr>
<td>jobBERT-de</td>
<td>general-domain data (12 GB)</td>
<td>in-domain data (3.7 GB)</td>
</tr>
<tr>
<td>GBERT</td>
<td>general-domain data (160 GB)</td>
<td>task data</td>
</tr>
<tr>
<td>jobGBERT</td>
<td>general-domain data (160 GB)</td>
<td>in-domain data (3.7 GB)</td>
</tr>
</tbody>
</table>

- 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:

<table>
<thead>
<tr>
<th>job ad text</th>
<th>profession</th>
<th>main task</th>
<th>experience</th>
</tr>
</thead>
<tbody>
<tr>
<td>... we are looking for an experienced pharmaceutical assistant as sales consultant ..</td>
<td>32 - Medical, pharmaceutical professions</td>
<td>7 - customer service, sales, cashier</td>
<td>1 - needed</td>
</tr>
</tbody>
</table>

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:

<table>
<thead>
<tr>
<th>Zone</th>
<th>Definition</th>
<th>rel. Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>z1</td>
<td>company description</td>
<td>17.2%</td>
</tr>
<tr>
<td>z2</td>
<td>reason of vacancy</td>
<td>0.5%</td>
</tr>
<tr>
<td>z3</td>
<td>administration &amp; residual text</td>
<td>25.2%</td>
</tr>
<tr>
<td>z4</td>
<td>job agency description</td>
<td>0.7%</td>
</tr>
<tr>
<td>z5</td>
<td>material incentives</td>
<td>1.7%</td>
</tr>
<tr>
<td>z6</td>
<td>job description</td>
<td>32.4%</td>
</tr>
<tr>
<td>z7</td>
<td>required hard skills</td>
<td>12.7%</td>
</tr>
<tr>
<td>z8</td>
<td>required personality (soft skills)</td>
<td>9.5%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>orig. train set 1990-2014 (n=22.5k), orig. test set 2010-2014 (n=650)</th>
<th>Balanced Accuracy</th>
<th>updated train set 1990-2021 (n=23.1k), new test set 2015-2021 (n=150)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT-de</td>
<td>0.855</td>
<td>0.826</td>
<td>0.847</td>
</tr>
<tr>
<td>jobBERT-de</td>
<td>0.874</td>
<td><strong>0.894</strong></td>
<td><strong>0.941</strong></td>
</tr>
<tr>
<td>GBERT</td>
<td>0.869</td>
<td>0.859</td>
<td>0.885</td>
</tr>
<tr>
<td>jobGBERT</td>
<td><strong>0.880</strong></td>
<td>0.876</td>
<td>0.896</td>
</tr>
</tbody>
</table>
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

