INFORMATION EXTRACTION FOR JOB MARKET ANALYSIS

Mike **Zhang**, Kristian **Nørgaard Jensen**, Barbara **Plank**
{mikz, krnj, bplank}@itu.dk

north / IT-UNIVERSITETET I KØBENHAVN

(1) De-identification of Privacy-related Entities in Job Postings [1]

…

We at [company] are committed to safeguarding and promoting the welfare of children.

In total, we have several locations in [location], [location], and [location].

Are you interested? Further information is available by contacting [profession] [name] on tel. [contact]. The application and relevant appendices are emailed to [profession] [name] at the email address [contact].

…

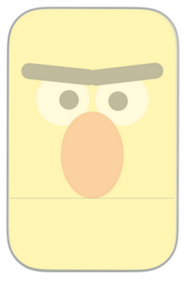
The need for de-identification technology is increasing, as privacy-preserving data handling is in high demand in many domains. In this work we focus on job postings.

We annotate for specific named entities:
`[Organization, Location, Profession, Contact, Name]`

We do this for 395 job postings from StackOverflow and release the data

Scan me!

We use the MaChAmp toolkit [2] with BERT [3] to predict the named entities and exploit auxiliary data via multi-task learning [4].



Model	Auxiliary Tasks	F1
$BERT_{base}$	JobStack	79.91±0.38
	JobStack + CoNLL	81.27±0.28
	JobStack + I2B2	**82.05±0.80**
	JobStack + CoNLL + I2B2	81.47±0.43

Evaluation of the best performing model on the test set across three runs, reported are the F1-score and standard deviation.

[1] Jensen, Kristian Nørgaard, Mike Zhang, and Barbara Plank. "De-identification of Privacy-related Entities in Job Postings." *NoDaLiDa 2021* (2021): 210.

[2] van der Goot, Rob, et al. "Massive Choice, Ample Tasks (MaChAmp): A Toolkit for Multi-task Learning in NLP." EACL: System Demonstrations. 2021.

[3] Devlin, Jacob, et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." NAACL: Human Language Technologies, Volume 1 (Long and Short Papers). 2019.

[4] Caruana, Rich. "Multitask learning." *Machine learning* 28.1 (1997): 41-75.

(2) Pilot on Skill Extraction from Job Postings

…

The successful candidate will be [self-motivated] and capable of [working on their own initiative], an excellent [communicator], with both customers and our highly motivated team at the company.

Experience in the [kitchen industry] and competent with [computers].

Furthermore, we are looking for someone who is a [team player] with a reliable approach to [problem solving].

[…] = skill
[…] = knowledge

…

Skill Extraction (SE) has been relatively well studied, but previous works frequently limit themselves to document-level SE and inferred meta-labels (i.e., a predefined skill inventory).

We instead annotate for token-level skills and knowledge components, following ESCO spans [5].

We have currently around ~400 job postings with around 12.5K annotated spans.

We use the MaChAmp toolkit with BERT, SpanBERT [6], and domain-adaptive pertained BERT & SpanBERT [7] to predict the skill and knowledge spans and compare single-task vs. Multi-task learning (i.e., predicting only skills/knowledge/both at the same time).

Model	Single-task F1	Multi-task F1
$BERT_{base}$	59.35±0.46	58.72±0.48
SpanBERT	58.69±0.36	58.88±0.28
JobBERT	**60.32±0.39**	59.44±0.81
JobSpanBERT	59.79±0.53	59.29±0.43

Evaluation of the best performing model on the development set across three runs, reported are the span F1-score and standard deviation.

[5] De Smedt, Johan, Martin Ie Vrang, and Agis Papantoniou. "ESCO: Towards a Semantic Web for the European Labor Market." *LDOW@ WWW*. 2015.

[6] Joshi, Mandar, et al. "SpanBERT: Improving Pre-training by Representing and Predicting Spans." *Transactions of the Association for Computational Linguistics* 8 (2020): 64-77.

[7] Gururangan, Suchin, et al. "Don't Stop Pretraining: Adapt Language Models to Domains and Tasks." *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020.