NNOSE: Nearest Neighbour Occupational Skill Extraction

Mike **Zhang**, Rob **van der Goot**, Min-Yen **Kan**, Barbara **Plank**

jjz@cs.aau.dk

IT UNIVERSITY OF CPH

Introduction & Motivation

- Skill Extraction (SE) is the task of extracting spans from job ads. Certain skills might be underrepresented in job description, resulting in a *sparsity of skills* SE datasets.
- In job descriptions, there is a *long-tail pattern,* popular skills are more commonly mentioned, while niche expertise appears less frequently.
- We explore Nearest Neighbor Language Models (NNLMs; Khandelwal et al., 2020), using the kNN algorithm as a retriever to retrieve context–token pairs from a datastore with LM encoders.

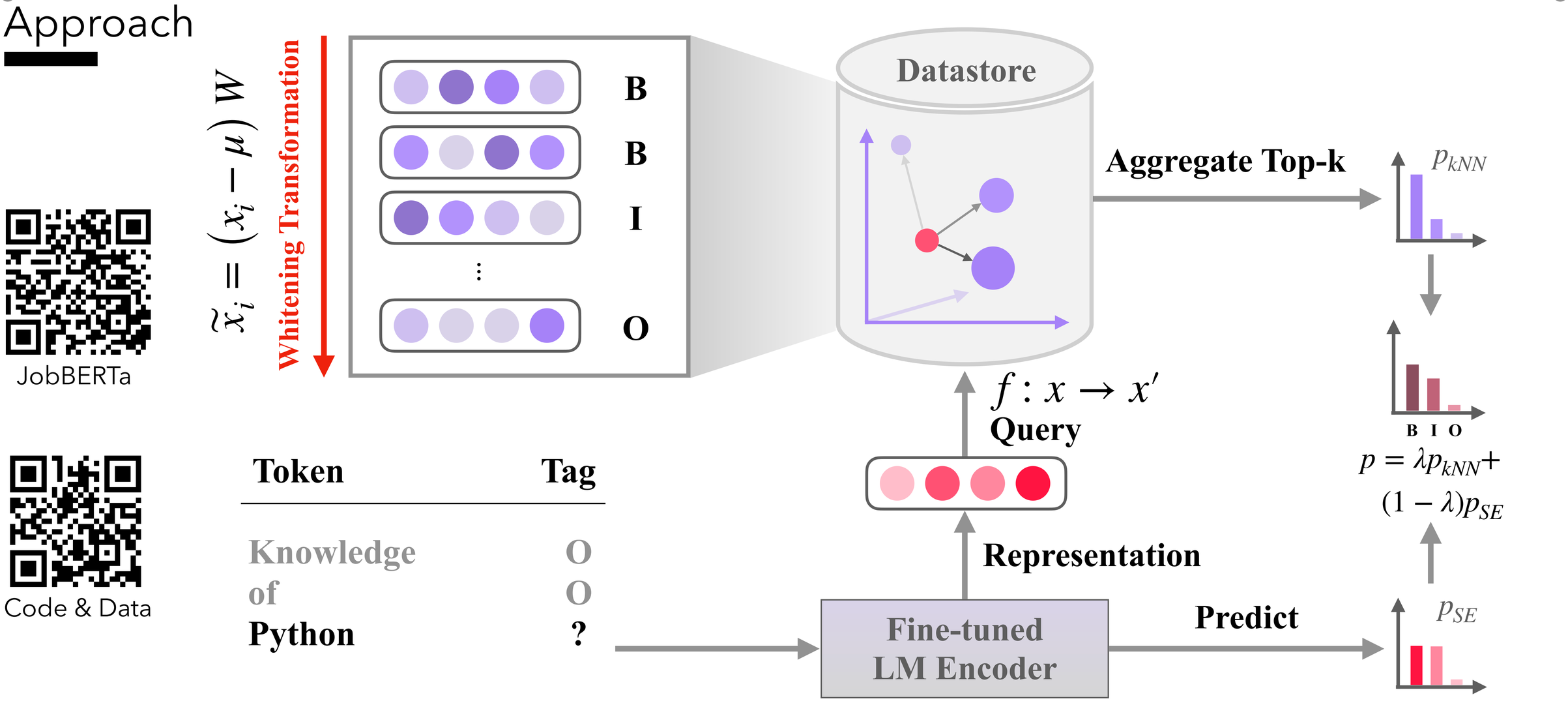
Models & Data

- JobBERT (Zhang et al., 2022)
- RoBERTa (Liu et al., 2019)
- JobBERTa (This work): RoBERTa further pre-trained on 3.2M job posting sentences.

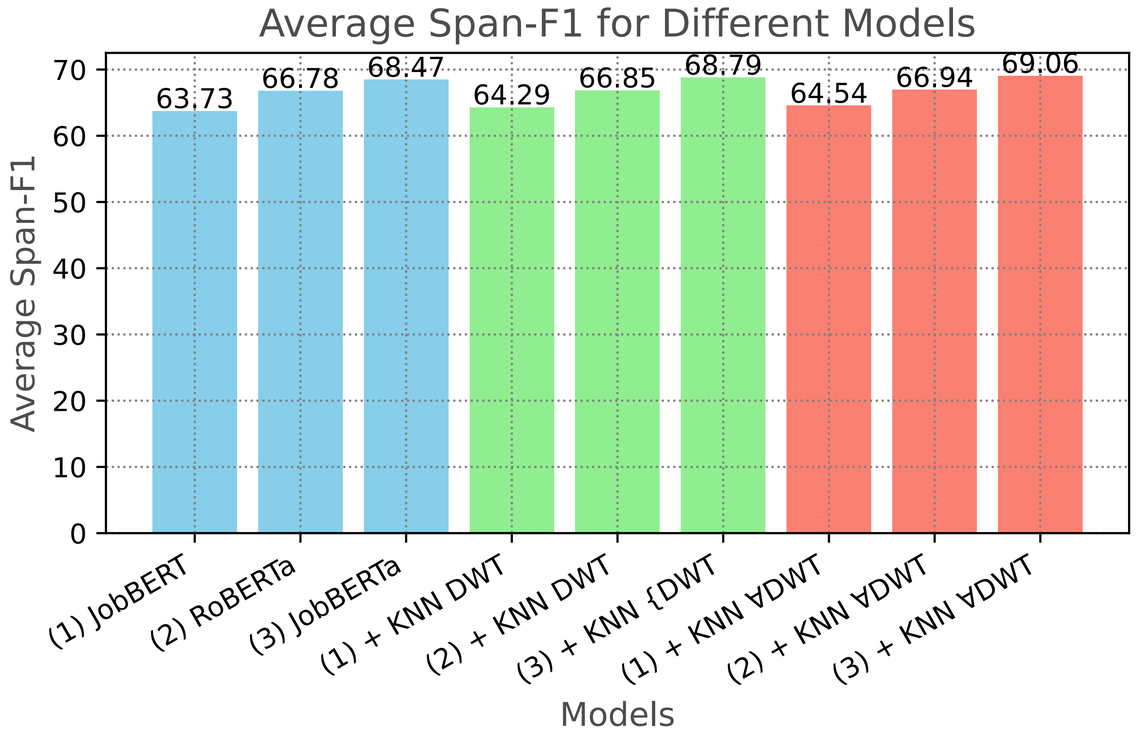
Dataset	Train	Dev	Test	D (tokens)
SkillSpan	5,866	3,992	4,680	86.5K
Sayfullina	3,706	1,854	1,853	53.1K
Green	8,670	963	336	209.5K
Total				348.2K

Table 1: **Dataset Statistics.** We provide statistics for all three datasets. Input granularity is at the token level, with performance measured in span-F1. The size of the datastore D is in tokens and determined by embedding tokens and their context from the training sets, resulting in approximately 350K keys.

Approach

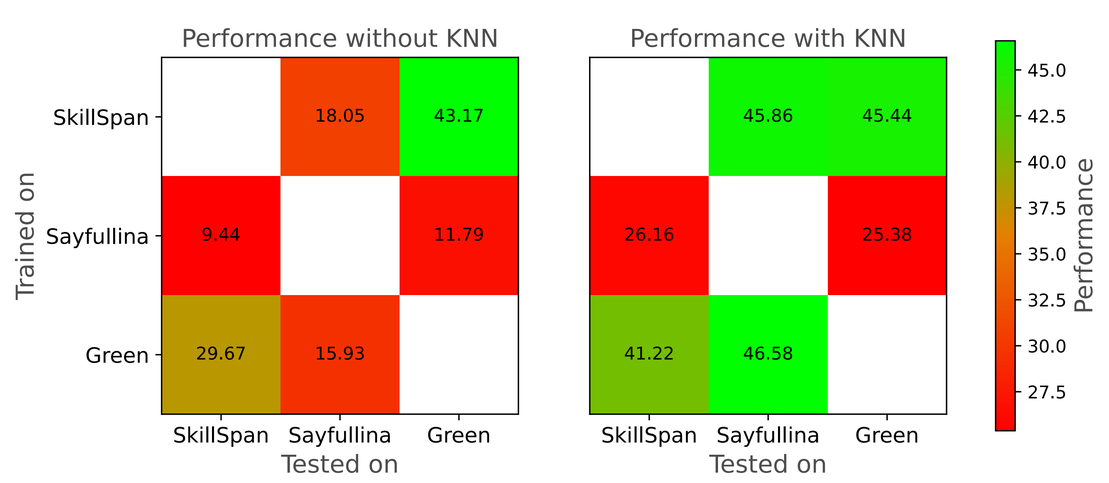


Results



- The best-performing baseline model is JobBERTa.
- All models seem to benefit from the NNOSE setup, JobBERT and JobBERTa shows the largest improvements, with the largest gains observed in the ∀D+WT datastore setup.

Analysis



- We observe large gains using NNOSE in a cross-dataset setting.
  - We train on one dataset and apply the model to another using the datastore.
- We confirm findings similar to Khandelwal et al. (2020), that memorisation using NNLMs improves recall.