Development of a Benchmark Corpus to Support Entity Recognition in Job Descriptions

Thomas AF Green, Diana Maynard, and Chenghua Lin

May 9, 2022

Background - Automatic Job Matching

- Online recruitment is becoming more common
- There is an overwhelming volume of data for individuals to review (information overload/filter failure)
- Automatic matching solutions need to replicate human decision-making
- Issue #1: CV/Job Description data is often unstructured
- Issue #2: Salient components of CV/JD are not clear
- Issue #3: It is difficult to evaluate systems
  - intrinsically - what is a good match?
  - extrinsically - lack of public data

Background - Skills as a Measure of Job Fit

- 'Skills' are used as indicator of job fit, however:
  - No academic consensus regarding definition
  - No public datasets of CVs/Job Descriptions with labelled Skills
- Developing a dataset of job descriptions with labelled Skills is challenging
  - Entities are 'fuzzy' - low inter-annotator agreement

Research Goal

Research Question

How can salient entities in applicant profiles and job descriptions be identified and extracted for use in an applicant profile-job description matching solution?

Scientific Contributions

1. A list of entity classifications and their definitions in the form of an annotation scheme for salient entities within applicant profiles and job descriptions, made publicly available
2. A public, labelled dataset for the development and evaluation of entity extraction systems
3. A state-of-the-art system for extracting salient entities from applicant profiles and job descriptions

Previous Approaches to Defining Skills

TAF Green
D Maynard
C Lin

A Skill is...

1. ...a term that exists in a section of a CV/Job Description titled 'Skills'
   - Maheshwari et al. (2010); Karakatsanis et al. (2017); Bastian et al. (2014); Kivimäki et al. (2020)
2. ...a term that exists in a Skills Database
   - O*NET[1]; ComputerHope[2]

[1] https://www.onetonline.org/
[2] https://www.computerhope.com/

Metrics

Agreement *(between annotators)*

- Pairwise F1-Score, calculated *without* the 'NONE' label
- Token-level Kappa, calculated *with* the 'NONE' label
- Token-level Kappa, calculated *without* the 'NONE' label

Accuracy *(vs. gold standard)*

- Pairwise F1-Score, calculated *without* the 'NONE' label

Labels

- **Skill**
  - e.g. ‘data analysis’, ‘French’, ‘honesty’
- **Qualification**
  - e.g. ‘Bachelor’s Degree’, ‘chartership’, ‘three A-levels’
- **Experience**
  - e.g. ‘2 years experience’, ‘minimum of 5 years experience’
- **Occupation**
  - e.g. ‘Teaching assistant’, ‘CEO’, ‘Chef de partie’
- **Domain**
  - e.g. ‘aerospace’, ‘oil industry’, ‘human resources’

Crowdsourcing Platform - Amazon Mechanical Turk

TAF Green
D Maynard
C Lin

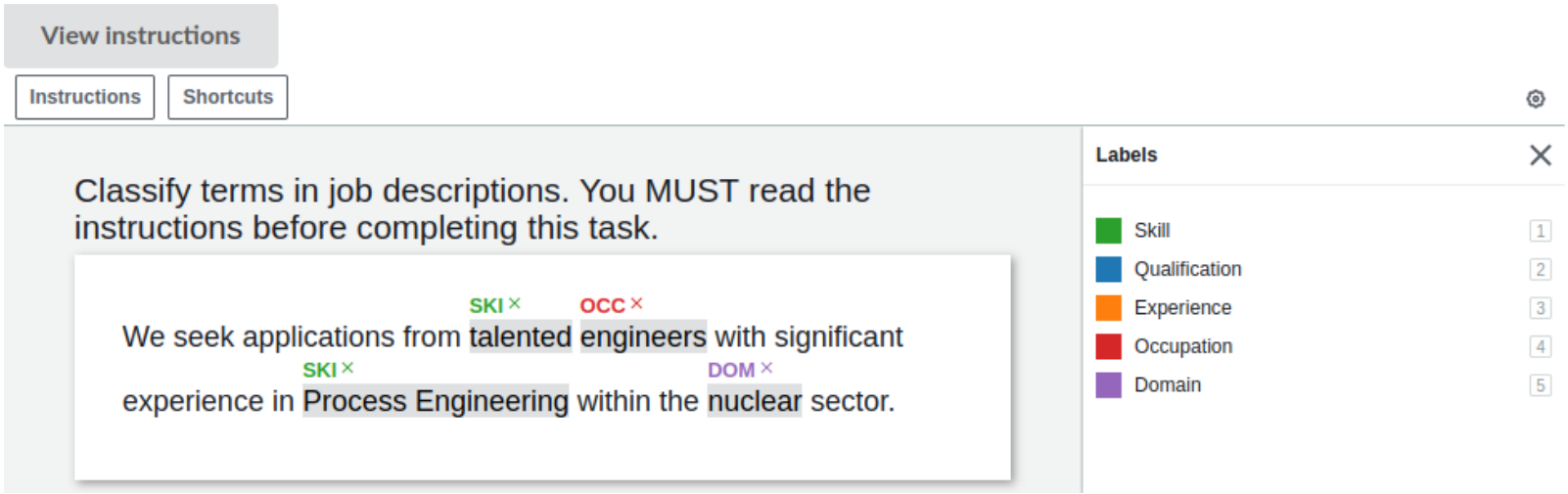
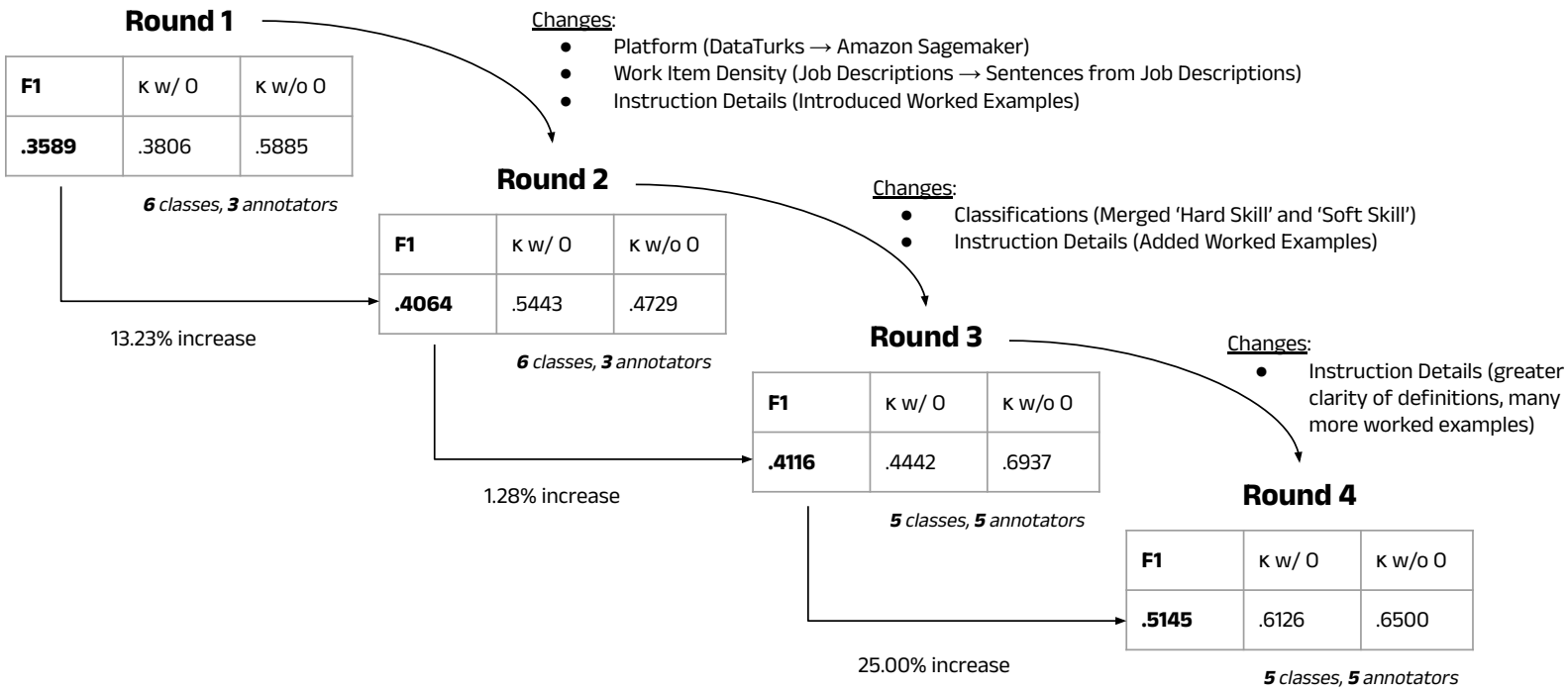


Figure 1: The GUI for the annotation task, hosted on AMT.

Developing the Annotation Schema

TAF Green
D Maynard
C Lin



Crowd Layer

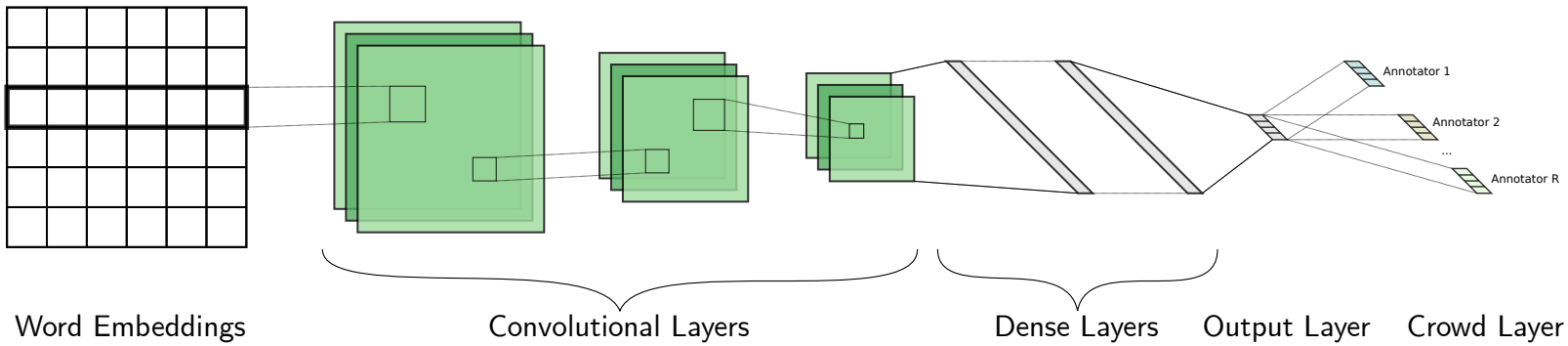


Figure 2: An example crowd layer appended to a CNN for NER with 4 classes and R annotators, adapted from Rodrigues and Pereira (2017).

Finding an Accuracy Threshold

1. Source a public, well-established, human-annotated NER dataset
2. Demonstrate NER model performance (using crowd layer)
3. Identify individual worker accuracies (F1 without O-label)
4. Rerun models with altered data across varying average accuracies
5. Investigate the relationship between worker accuracy and model performance

Model Architecture

- Using Rodrigues and Pereira (2017) implementation of the Crowd Layer
- Model:
  - 300d word embeddings layer
    - GloVe 6B Wikipedia + Gigaword 5
  - 5x5 convolutional layer with 512 features
  - GRU cell with 50d hidden state
  - Fully connected layer with softmax activation
  - Categorical cross-entropy loss used, with Adam optimizer

Ground Truth Labels

Testing Accuracy	Precision	Recall	F1
.95	.7733	.7292	.7506

trained over 20 epochs

Worker Labels with Majority Voting

Testing Accuracy	Precision	Recall	F1
.91	.6761	.4460	.5375

trained over 10 epochs

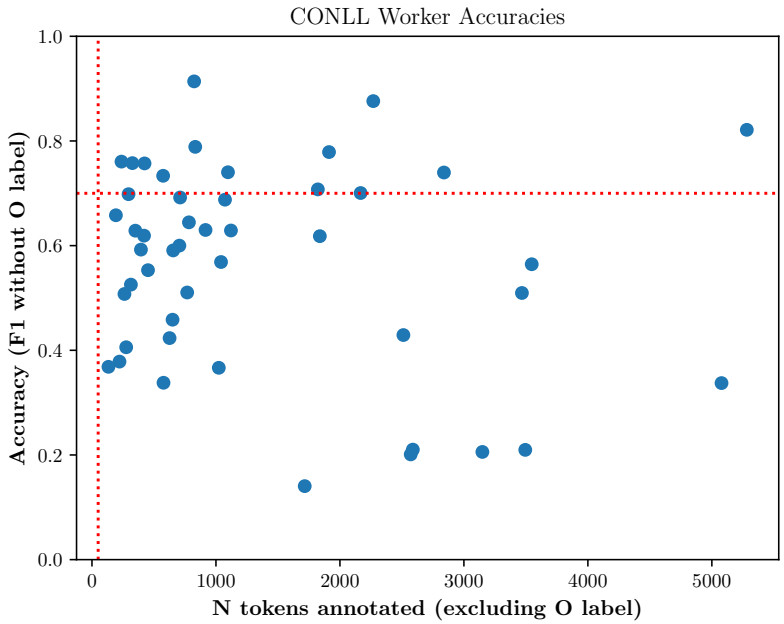
Worker Labels with Crowd Layer

Testing Accuracy	Precision	Recall	F1
.93	.6698	.5669	.6141

pre-trained with MV data for 5 epochs
trained over 30 epochs

2003 CONLL NER Task

- Four types of Named Entities:
  - **Persons**
  - **Locations**
  - **Organisations**
  - **Misc**
- 5,985 Sentences
- Crowdsourced annotations from Amazon Mechanical Turk, **47** annotators (2 per sentence)

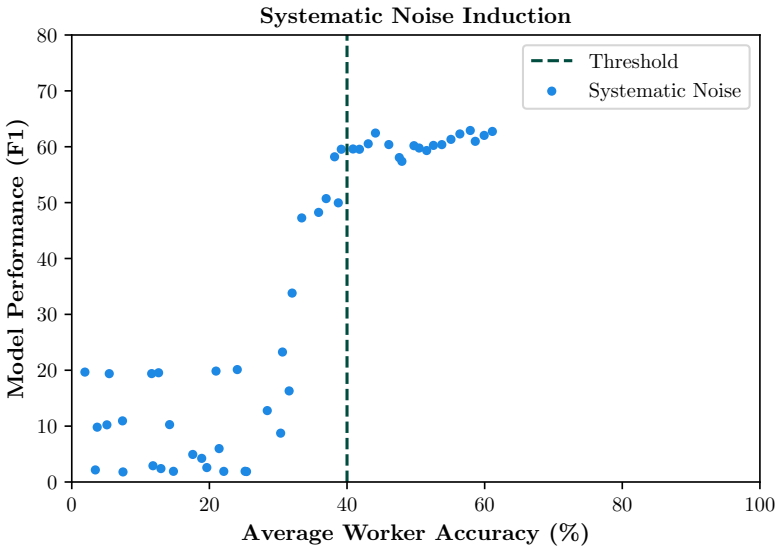


The Relationship Between Worker Accuracy and Model Performance

TAF Green
D Maynard
C Lin

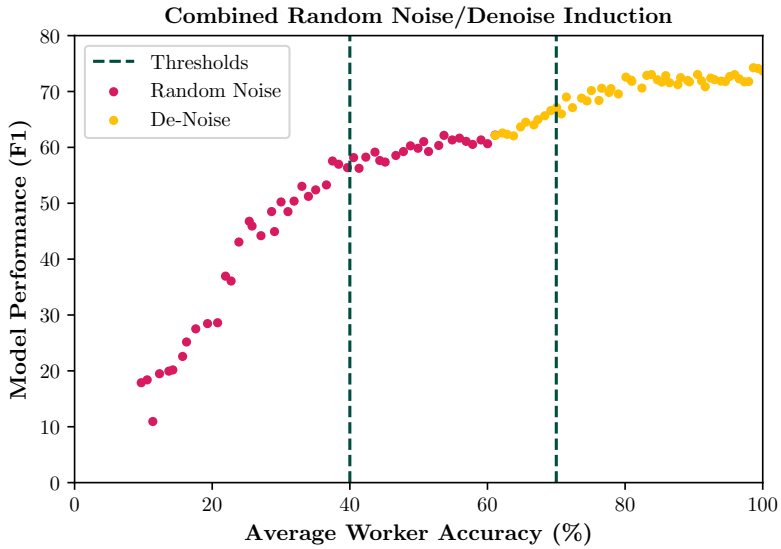
Systematic Noise

Annotators make consistent errors; when they are incorrect, it is because they are consistently misclassifying class A as class B



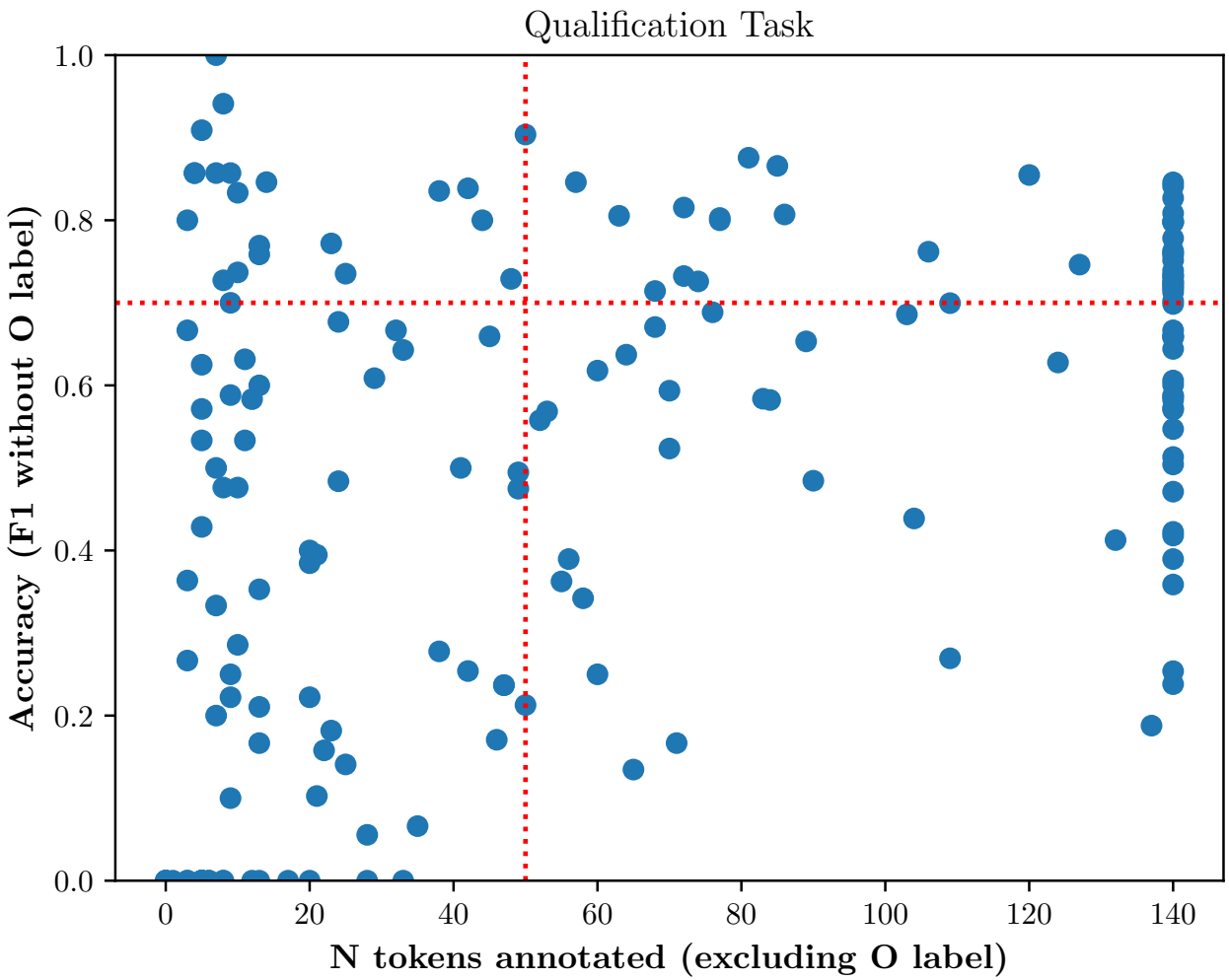
Random Noise

Annotators make random errors; when they are incorrect, each other (incorrect) classification is equally likely to be selected



AMT Results

TAF Green
D Maynard
C Lin



Corpus Statistics

Sentences	10,000
Tokens	245,606
Avg. tokens per sentence	24.6
Annotation spans (post aggregation)	18,617
Annotated tokens (post aggregation)	79,826
Avg. tokens per annotation	4.3
Number of independent Annotators	25

Table 1: Annotated corpus statistics.

Corpus Statistics: Entity Class Distribution

Label	Frequency	Proportion
Skill	66,732	28.56%
Occupation	6,117	2.62%
Domain	3,705	1.59%
Experience	1,328	0.57%
Qualification	1,944	0.83%
None	153,802	65.83%
Total	**233,628**	

Table 2: Class distribution for the live, aggregated corpus (one label per token).

Corpus Statistics: Inter-Annotator Agreement

Cohen's κ on all tokens	0.49
Cohen's κ on annotated tokens only	0.73
Krippendorff's α	0.55
F_1 on annotated tokens only	0.90

Table 3: IAA on the live corpus, calculated by averaging pairwise comparisons between all combinations of annotators where both annotators labelled a shared item.

Data Preprocessing

- **Label Aggregation**
  - For non-Crowd Layer model training
  - Using Worker Qualification Score
- **Reclassification of 'Experience' Spans to 'Skill'**
  - Correcting cases with no 'time' element, e.g. '*experience managing clients*'
- **Splitting Multi-Term Spans**
  - e.g. splitting the annotated span '*Asbestos Surveyors, Lead Asbestos Surveyors*' into two distinct entities bounded by a comma

Baseline Model: Conditional Random Fields

- Uses NLTK method of feature preparation (morphological features of target and adjacent tokens)
- Optimised L1 and L2 regularization coefficients found via Randomized Search

Label	P	R	F_1	Support
B-Skill	0.69	0.37	0.48	676
I-Skill	0.53	0.71	0.61	1429
B-Qualification	0.72	0.50	0.59	26
I-Qualification	0.39	0.23	0.29	40
B-Occupation	0.90	0.65	0.75	137
I-Occupation	0.93	0.71	0.81	164
B-Experience	0.86	0.67	0.75	9
I-Experience	0.42	0.76	0.54	17
B-Domain	0.53	0.40	0.46	60
I-Domain	0.34	0.28	0.31	39
micro avg	0.58	0.60	0.59	2597
macro avg	0.63	0.53	0.56	2597
weighted avg	0.61	0.60	0.58	2597

Table 4: Results for CRF model (trained on preprocessed data). Precision, Recall, and F_1-Score are presented.

Further Models

Model	**Precision**	**Recall**	F_1
BERT, base, cased	0.70	0.75	0.73
BERT, base, multilingual, cased	0.72	0.72	0.72
BERT, base, uncased	0.63	0.67	0.65
CNN + CrowdLayer*	0.71	0.55	0.62
CRF (Baseline)	*0.58*	*0.60*	*0.59*
DistilBERT, base, uncased	0.55	0.55	0.55
BiLSTM-CRF	0.53	0.45	0.49
ALBERT, base	0.48	0.50	0.49

Table 5: Micro-Averaged results for additional ER models trained on the published dataset.

*Uses raw labels (as opposed to aggregated labels)

Conclusion

- Developing a dataset of salient entities in job descriptions is challenging because annotator disagreement is likely to be high on fuzzy entities
- Our proposed solution for crowdsourcing data on an ambiguous annotation task combines:
  1. An iterative approach to guideline and task development
  2. A qualification task to ensure Workers achieve an acceptable score (set at 0.7) on the task before being allowed to contribute

Thank You

This work was supported by the Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications funded by UK Research and Innovation [grant number EP/S023062/1].

This work was also supported by TribePad.

References I

M. Bastian, M. Hayes, W. Vaughan, S. Shah, P. Skomoroch, and H. Kim. Linked in skills: Large-scale topic extraction and inference. *RecSys 2014 - Proceedings of the 8th ACM Conference on Recommender Systems*, (October):1–8, 2014. doi: 10.1145/2645710.2645729.

I. Karakatsanis, W. AlKhader, F. MacCrory, A. Alibasic, M. A. Omar, Z. Aung, and W. L. Woon. Data mining approach to monitoring the requirements of the job market: A case study. *Information Systems*, 65:1–6, apr 2017. ISSN 03064379. doi: 10.1016/j.is.2016.10.009.

I. Kivimäki, A. Panchenko, A. Dessy, D. Verdegem, P. Francq, C. Fairon, H. Bersini, and M. Saerens. A graph-based approach to skill extraction from text. *Proceedings of TextGraphs@EMNLP 2013: The 8th Workshop on Graph-Based Methods for Natural Language Processing*, (October):79–87, 2020.

References II

S. Maheshwari, S. Abhishek, and P. K. Reddy. An Approach to Extract Special Skills to Improve the Performance of Resume Selection. *Conference: Databases in Networked Information Systems, 6th International Workshop, DNIS 2010*, (March 2010), 2010. ISSN 03029743. doi: 10.1007/978-3-642-12038-1. URL http://www.scopus.com/inward/record.url?eid=2-s2.0-77951565566{&}partnerID=tZOtx3y1.

F. Rodrigues and F. Pereira. Deep learning from crowds. (July), sep 2017. URL http://arxiv.org/abs/1709.01779.