DATASHEET:
JOB DESCRIPTIONS ENTITY RECOGNITION CORPUS

Anonymous ACL-IJCNLP submission

This document is based on *Datasheets for Datasets* by Gebru *et al.* [1].

MOTIVATION

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

The identification and extraction of salient entities is an important task in many real-world information extraction applications such as text classification, efficient search algorithms, and content recommendations. In areas such as recruitment, job-seekers and recruiting companies alike benefit from systems that automatically and continuously acquire up-to-date information about listed job roles and applicant profiles in terms of the skills, qualifications, and experience they have or require.

However, in the recruitment domain, the development of Entity Recognition (ER) models to perform these tasks is severely hindered by the lack of publicly available training data. Many available ER corpora consist of general news articles [2], while information about job descriptions is typically only available on online job portals.

The purpose of this dataset is to establish a standard definition of the relevant entities in order to make their extraction from unstructured job descriptions easier and more reliable.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

Details removed for anonymous review.

What support was needed to make this dataset? (e.g.who funded the creation of the dataset? If there is an associated grant, provide the name of the grantor and the grant name and number, or if it was supported by a company or government agency, give those details.)

Details removed for anonymous review.

Any other comments?

COMPOSITION

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?

```
Support ? ? ? B-Skill B-Skill ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ?
the ? ? ? I-Skill I-Skill ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ?
Sourcing ? ? ? I-Skill I-Skill ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ?
Teams ? ? ? I-Skill I-Skill ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ?
on ? ? ? I-Skill O ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ?
future ? ? ? I-Skill O ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ?
fabric ? ? ? I-Skill O ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ?
requirements ? ? ? I-Skill O ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ?
and ? ? ? I-Skill O ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ?
innovations ? ? ? I-Skill O ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ?
. ? ? ? O O ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ?
```

Fig. 1. An example instance taken from the file `raw_data/answers.txt`.

Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

Each instance is a sentence taken from a job description, with each token in the sentence assigned 24 labels; one per human annotator. An example instance is shown in Figure 1.

How many instances are there in total (of each type, if appropriate)?

There are 10,000 instances in the dataset with a total of 245,606 tokens. Additionally, a 'test set' is provided, which includes an independent 586 items.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

10,000 sentences from job descriptions were sampled at random from a larger dataset of job descriptions. All of the sampled instances were annotated and are included in the dataset.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Each instance contains a variable number of tokenised words and punctuation.

Is there a label or target associated with each instance? If so, please provide a description.

Each token is assigned 24 labels; one per annotator. Labels are one of five distinct classes (Skill, Experience, Occupation, Qualification, Domain), a 'None' label (O), or a '?' denoting that annotator did not annotate that token. Class labels include their BIO prefix, denoting that label begins a span ('B'), is inside a span ('I'), or is outside, i.e. not part of a span ('O').

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

HTML tags and non-unicode characters were removed from instances prior to annotation.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

There are no relationships between individual instances.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

A separate test set is provided, containing 586 instances with one gold standard label associated with each token.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

There are no sources of noise other than that associated with crowdsourced annotations; agreement between annotators is not perfect.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset is completely self-contained.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.

There is no confidential or non-public communication information included in the dataset.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

No.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

Although labels were crowdsourced by human annotators, the data itself does not relate to people.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

The raw form of the dataset contains public email addresses of recruiting professionals and HR personnel.

Any other comments?

COLLECTION

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

Instances were directly observable as raw text and the labels were extracted by the process described below. Job description data was collected by downloading a public dataset hosted on Kaggle[1].

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data

[1] https://www.kaggle.com/airiddha/trainrev1

associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created. Finally, list when the dataset was first published.
The original dataset of job descriptions was published September 2018. Annotations were sourced and published in July 2021.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?
Human Workers on the Amazon Mechanical Turk[2] platform were paid to annotate instances of job descriptions. In order to qualify to contribute to the dataset, Workers were required to pass a test (20 sampled instances where the gold standard was known) with at least 70% accuracy.

What was the resource cost of collecting the data? (e.g. what were the required computational resources, and the associated financial costs, and energy consumption - estimate the carbon footprint)
The total cost of qualification and live data collection was $2,139.64.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

Workers on the Amazon Mechanical Turk platform annotated instances. Workers were compensated at $0.04/HIT for the 20-item qualification task and $0.08/HIT for the live task, the latter equating to the standard minimum wage in the country in which the task was deployed. These figures were calculated using the average time that it took for an instance to be annotated during task development.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.
A formal ethics application was made to the affiliated university's ethics application system. This was approved in August 2020.

Does the dataset relate to people? If not, you may skip the remainder of the questions in this section.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate)

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis)been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

Any other comments?

PREPROCESSING / CLEANING / LABELING

Was any preprocessing/cleaning/labeling of the data done(e.g.,discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.
Instances were tokenised to show 24 labels for each token (one label per annotator). Additionally, class labels were BIO tagged.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.
Original job description data continues to be hosted on

[2] https://www.mturk.com/

Kaggle[3].

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.
Labelling through Amazon Mechanical Turk used a bespoke interface included in the published repository along with Python scripts for data preprocessing and label aggregation.

Any other comments?

USES

Has the dataset been used for any tasks already? If so, please provide a description.
A benchmark entity recognition model was trained using the dataset and is detailed in the associated research paper.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

What (other) tasks could the dataset be used for?
The primary use of this data is for training Entity Recognition systems. Although this dataset could also be used for text analysis or for training unsupervised models, the original dataset of job descriptions without human labels, of which this data is a subset, may be more suitable.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

Are there tasks for which the dataset should not be used? If so, please provide a description.

Any other comments?

DISTRIBUTION

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.
The dataset is made publicly available.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?
The dataset is published as a public GitHub repository.

When will the dataset be distributed?
The dataset will be distributed formally pending anonymous review of the accompanying research paper.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.
The dataset is distributed under a Creative Commons license (CC-BY).

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.
There are no known restrictions on the data associated with the instances.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

Any other comments?

MAINTENANCE

Who is supporting/hosting/maintaining the dataset?
The dataset is hosted as a public repository on GitHub. Issues raised through this platform will be addressed and maintained by the first author of the associated research paper.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?
The author of the dataset can be contacted by email (redacted for anonymous review).

Is there an erratum? If so, please provide a link or other access point.

[3] https://www.kaggle.com/airiddha/trainrev1

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?
There are no plans to update the dataset following publication.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.
The dataset can be contributed to using GitHub's 'pull request' features, which will be monitored and reviewed by the first author.

Any other comments?

References

[1] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé, and Kate Crawford. Datasheets for Datasets. 2018.

[2] Nolan Lawson and Kevin Eustice. Annotating Large Email Datasets for Named Entity Recognition with Mechanical Turk. *Computational Linguistics*, (June):71–79, 2010.